TRƯỜNG ĐẠI HỌC QUY NHƠN

Bài giảng

NGHIÊN CỨU ĐỊNH LƯỢNG TRONG KẾ TOÁN-KIỂM TOÁN

GV: TS. Trương Thị Thanh Phượng

Chương 1

TỔNG QUAN VỀ NGHIÊN CỨU KẾ TOÁN-KIỂM TOÁN

1.1. Khái niệm nghiên cứu và cách tiếp cận nghiên cứu

1.1.1. Khái niệm nghiên cứu

Nghiên cứu là sự tìm kiếm kiến thức, hoặc là sự điều tra mang tính hệ thống, với suy nghĩ mở rộng để khám phá, giải thích và phát triển các phương pháp nhằm vào sự tiến bộ kiến thức của nhân loại.

Theo Babbie (1986) : Nghiên cứu là quá trình thu thập dữ liệu và phân tích thông tin một cách hệ thống nhằm tăng cường sự hiểu biết về một hiện tượng.

Theo Kothari (2004) : Nghiên cứu là quá trình thu thập, phân tích dữ liệu một cách có hệ thống nhằm khám phá các vấn đề liên quan.

Theo Kumar (2014) : Nghiên cứu là một trong những cách tìm ra các câu trả lời cho các câu hỏi.

Theo Shuttleworth (2008) : Nghiên cứu là bao gồm mọi cách thức thu thập dữ liệu, thông tin và sự kiện cho sự phát triển kiến thức.

Nghiên cứu là quá trình thu thập và phân tích thông tin một cách có hệ thống nhằm tăng cường sự hiểu biết về một hiện tượng hay một vấn đề nào đó.

Nghiên cứu khoa học

Theo Babbie (2011) : Nghiên cứu khoa học là cách thức : con người tìm hiểu các hiện tượng khoa học một cách có hệ thống và quá trình áp dụng các ý tưởng, nguyên lý để tìm ra các kiến thức mới nhằm giải thích các sự vật hiện tượng.

Theo Armstrong và Sperry (1994) : Nghiên cứu khoa học dựa vào việc ứng dụng các phương pháp khoa học để phát hiện ra những cái mới về bản chất sự vật, về thế giới tự nhiên và xã hội, và để sáng tạo phương pháp và phương tiện kỹ thuật mới cao hơn, giá trị hơn. Hình thức nghiên cứu này cung cấp thông tin và lý thuyết khoa học nhằm giải thích bản chất và tính chất của thế giới.

Nghiên cứu khoa học là hoạt động tìm kiếm, xem xét, điều tra, hoặc thử nghiệm. Dựa trên những số liệu, tài liệu, kiến thức,… đạt được từ các thí nghiệm nghiên cứu khoa học để phát hiện những cái mới về bản chất sự vật, về thế giới tự nhiên và xã hội, và để sáng tạo phương pháp và phương tiện kỹ thuật mới cao hơn, giá trị hơn. Con người muốn nghiên cứu khoa học phải có kiến thức nhất định về lĩnh vực nghiên cứu và cái chính là phải rèn luyện cách làm việc tự lực, có phương pháp từ lúc ngồi trên ghế nhà trường.

Vai trò của nghiên cứu khoa học

Nghiên cứu là để truyền tải thông tin. Tuy nhiên một bài nghiên cứu hiệu quả phải:

- Làm thay đổi cách nhìn nhận vấn đề của người đọc.
- Thuyết phục người đọc tin vào một điều gì đó.
- Đưa người đọc đến quyết định và hành động.
- Dẫn dắt người đọc theo một quy trình nào đó.

Phương pháp nghiên cứu

Theo Yang (2001), phương pháp nghiên cứu cung cấp các chi tiết của quy trình và phương pháp cụ thể để thực hiện một vấn đề nghiên cứu. Phương pháp nghiên cứu cung cấp các quy trình cụ thể và chi tiết làm thế nào để bắt đầu, thực hiện và hoàn thành nhiệm vụ nghiên cứu và chủ yếu là tập trung vào làm thế nào để thực hiện được nghiên cứu.

Theo nghĩa hẹp, phương pháp nghiên cứu là phương pháp thu thập thông tin. Connaway và Powell (2003) cho rằng có rất nhiều cách để có được thông tin. Các phương pháp nghiên cứu phổ biến nhất là tìm kiếm tài liệu, hội thảo, hội thảo nhóm, phỏng vấn cá nhân, các cuộc điều tra qua điện thoại, các cuộc điều tra qua thư bưu điện và điều tra qua thư điện tử và mạng.

Phương pháp nghiên cứu đối với nhà kinh tế là tìm hiểu bản chất những vấn đề kinh tế đang cần giải quyết của hộ gia đình, doanh nghiệp, cộng đồng hay các nhà hoạch định chính sách ở phạm vi địa phương, quốc gia hay cộng đồng quốc tế nói chung. Các nhà kinh tế đã phát triển lý thuyết về cách thức hoạt động của thị trường, làm thế nào các hoạt động kinh tế được tiến hành trong các quốc gia hoặc trên phạm vi toàn cầu. Do đó, nghiên cứu kinh tế được chia thành hai nhóm chính: Nghiên cứu cơ bản và nghiên cứu ứng dụng.

1.1.2. Cách tiếp cận nghiên cứu

Giới thiệu về phương pháp nghiên cứu định tính

Nghiên cứu định tính là một dạng nghiên cứu được tạo thành từ các phương pháp kỹ thuật chuyên môn được sử dụng để tìm hiểu sâu về các phản ứng từ trong suy nghĩ và tình cảm của con người

Theo Marshall và Rossman (1998): Nghiên cứu định tính là một phương pháp điều tra được sử dụng trong nhiều lĩnh vực khoa học kỹ thuật khác nhau, không những trong khoa học xã hội mà còn trong nghiên cứu kinh tế. Mục đích của nghiên cứu định tính nhằm tìm kiếm sự hiểu biết sâu sắc về hành vi của con người và lý do chi phối hành vi như vậy. Nghiên cứu định tính điều tra lý do tại sao và làm như thế nào của việc ra quyết định, không chỉ là những gì, mà còn ở đâu, khi nào. Nghiên cứu định tính đưa ra các kết luận tổng quát hơn là các kết luận cụ thể:

Nghiên cứu định tính là loại hình nghiên cứu nhằm mô tả sự vật hiện tượng mà không quan tâm đến sự biến thiên của đối tượng nghiên cứu và không nhằm lượng hóa sự biến thiên này.

Nghiên cứu định tính hướng đến ý nghĩa các khái niệm, định nghĩa, đặc điểm và sự mô tả đối tượng nghiên cứu.

Nghiên cứu định tính thường được áp dụng trong xã hội học, tâm lý học, kinh tế học, kinh tế chính trị, luật,…

Nghiên cứu định tính thường dùng các phương pháp phân tích, so sánh, tổng hợp, mô tả, logic,…

Nghiên cứu định tính thường được áp dụng giai đoạn thu thập dữ liệu và phân tích dữ liệu. Ở giai đoạn thu thập dữ liệu, các kỹ thuật nghiên cứu định tính thường được áp dụng gồm nghiên cứu lý thuyết nền, phỏng vấn nhóm, phỏng vấn chuyên gia, nghiên cứu tình huống, quan sát,…

Ở giai đoạn phân tích dữ liệu, nghiên cứu định tính được sử dụng các kỹ thuật phân tích nội dung với các dữ liệu thu thập, quan sát hành vi cũng như các chứng cứ, sự kiện thu thập được.

Giới thiệu về phương pháp nghiên cứu định lượng

Theo Ehrenberg (1994): Nghiên cứu định lượng thường được sử dụng để kiểm định lý thuyết dựa vào cách tiếp cận suy diễn.

Theo Daniel Muijs, (2004) : Nghiên cứu định lượng là phương pháp giải thích hiện tượng thông qua phân tích thống kê với dữ liệu định lượng thu thập được.

Nghiên cứu định lượng là loại hình nghiên cứu mà ta muốn lượng hóa sự biến thiên của đối tượng nghiên cứu và công cụ thống kê, mô hình hóa được sử dụng cho việc lượng hóa các thông tin của nghiên cứu định lượng.

Các phương pháp định lượng bao gồm các quy trình thu thập dữ liệu, phân tích dữ liệu, giải thích và viết kết quả nghiên cứu.

Các phương pháp này liên quan đến sự xác định mẫu, chiến lược điều tra, thu thập dữ liệu, phân tích dữ liệu, trình bày kết quả nghiên cứu, thảo luận kết quả và viết công trình nghiên cứu.

Nghiên cứu định lượng phù hợp với các nghiên cứu xác định các yếu tố ảnh hưởng một kết quả nào đó. Cách tiếp cận định lượng thực hiện khi cần kiểm định các giả thuyết khác nhau và một lý thuyết nào đó.

So sánh phương pháp nghiên cứu định tính, định lượng

Thứ tự	Nội dung	Định tính	Định lượng
1	Mục tiêu Nghiên cứu	Hiểu sâu sắc, xây dựng lý thuyết	Mô tả hoặc dự báo, xây dựng hoặc kiểm định lý thuyết

2	Thiết kế nghiên	Có thể điều chỉnh trong quá trình thực hiện.	Được quyết định trước khi bắt đầu nghiên cứu.
	cứu	Thường phối hợp nhiều phương pháp	Sử dụng một hay phối hợp nhiều phương pháp.
3	Chọn mẫu, cỡ mẫu	Phi xác suất, có mục đích. Cỡ mẫu nhỏ	Xác suất, cỡ mẫu lớn
4	Phân tích dữ liệu	Phân tích bằng con người và thực hiện liên tục trong quá trình nghiên cứu	Phân tích bằng máy. Các phương pháp toán và thống kê làm chủ đạo. Phân tích có thể diễn ra suốt quá trình nghiên cứu.

Bảng 1.1. So sánh nghiên cứu định tính và nghiên cứu định lượng.

Giới thiệu phương pháp nghiên cứu hỗn hợp

Nghiên cứu phối hợp giữa định tính và định lượng được sử dụng khá phổ biến trong các ngành kinh tế, quản trị, tài chính,…

Trong quá trình nghiên cứu, chúng ta muốn hiểu rõ bản chất sự vật, nghiên cứu cơ sở lý thuyết, xây dựng mô hình hay khung phân tích thì phải dùng phương pháp nghiên cứu định tính với các công cụ tổng hợp, phân tích, so sánh đối chiếu, chuyên gia,…

Đồng thời, trong nghiên cứu chúng ta thường dựa trên một quan sát với cỡ mẫu đủ lớn để có kết quả tin cậy cần thiết. Chúng ta dùng dữ liệu, thông tin của mẫu để ước đoán số liệu, thông tin tổng thể nghiên cứu. Vì vậy phương pháp định lượng là hiển nhiên.

Khi so sánh nghiên cứu định lượng và định tính, Wilson (1982) lập luận rằng, việc sử dụng cân bằng cả hai nên được sử dụng trong nghiên cứu hiện đại. Thực hiện sự kết hợp này là sử dụng thế mạnh của cả hai nghiên cứu định tính và định lượng. Hơn nữa, phương pháp hỗn hợp có thể giải quyết sự phức tạp ngày càng tăng của thực tế xã hội. Hiểu biết có được sự kết hợp của cả hai phương pháp nghiên cứu định tính và định lượng cung cấp một sự hiểu biết tốt hơn và mở rộng của chủ đề nghiên cứu. Ta có bảng kết quả so sánh về quy trình các phương pháp như sau:

Các phương pháp nghiên cứu định tính	**Các phương pháp nghiên cứu định lượng**	**Các phương pháp nghiên cứu hỗn hợp**
Các phương pháp mới	Các câu hỏi dựa vào một	Cả hai phương pháp mới

nổi. Các câu hỏi mở. Dữ liệu phỏng vấn, dữ liệu quan sát, dữ liệu văn bản, và dữ liệu nghe nhìn.	công cụ xác định trước. Dữ liệu về kết quả hoạt động, dữ liệu về thái độ, dữ liệu quan sát, và dữ liệu tổng điều tra thống kê. Phân tích thống kê.	nổi và xác định trước. Cả câu hỏi có mở và đóng. Nhiều hình thức thu thập dữ liệu từ mọi khả năng. Phân tích thống kê và văn bản.

Bảng 1.2. So sánh quy trình phương pháp định tính, định lượng và hỗn hợp.

1.2. Phương pháp nghiên cứu

Phương pháp nghiên cứu khoa học là cách thức, con đường, phương tiện để giải quyết các nhiệm vụ nghiên cứu nhằm đạt được mục đích nghiên cứu. Dưới góc độ thông tin: phương pháp nghiên cứu khoa học là cách thức, con đường, phương tiện thu thập, xử lý thông tin khoa học (số liệu, sự kiện) nhằm làm sáng tỏ vấn đề nghiên cứu để giải quyết nhiệm vụ nghiên cứu và cuối cùng đạt được mục đích nghiên cứu. Nói cách khác, phương pháp nghiên cứu khoa học là những phương thức thu thập và xử lý thông tin khoa học nhằm mục đích thiết lập những mối liên hệ và quan hệ phụ thuộc có tính quy luật và xây dựng lý luận khoa học mới. Dưới góc độ hoạt động: phương pháp nghiên cứu khoa học là hoạt động có đối tượng, chủ thể (người nghiên cứu) sử dụng những thủ thuật, biện pháp, thao tác tác động, khám phá đối tượng nghiên cứu nhằm làm biến đổi đối tượng theo mục tiêu mà chủ thể tự giác đặt ra để thỏa mãn nhu cầu nghiên cứu của bản thân.

1.3. Quy trình nghiên cứu

1.3.1. Khái niệm quy trình nghiên cứu

Theo Kumar (2005), Quy trình nghiên cứu là một chuỗi các hành động diễn ra theo trình tự và gắn liền với nền tảng kiến thức cũng như các bước tư duy logic. Trong khái niệm này, quy trình nghiên cứu bao gồm một chuỗi các bước tư duy và vận dụng kiến thức về phương pháp nghiên cứu, kiến thức chuyên ngành, khởi đầu từ đặt vấn đề cho đến khi tìm ra câu trả lời. Các bước trong quy trình nghiên cứu phải theo một trình tự nhất định.

1.3.2. Nội dung các bước của quy trình nghiên cứu

Xác định và mô tả vấn đề nghiên cứu

a. Xác định vấn đề nghiên cứu

Đây là bước khá quan trọng trong quá trình nghiên cứu vì nếu không xác định đúng đề tài nghiên cứu sẽ khó thực hiện được. Để xác định vấn đề nghiên cứu ta cần làm rõ từng bước sau:

Bước 1. Xác định rõ lĩnh vực nghiên cứu.

Bước 2. Xác định loại vấn đề nghiên cứu.

Bước 3. Xác định sự cần thiết của nghiên cứu.

Bước 4. Đánh giá tính khả thi của nghiên cứu.

Bước 5. Trao đổi với giáo viên hướng dẫn, bạn bè, chuyên gia trong lĩnh vực này.

Bước 6. Thỏa mãn sự ưu thích, đam mê và sở trường của người nghiên cứu.

Ví dụ 1. Xác định vấn đề nghiên cứu. Đề tài: " Nghiên cứu các yếu tố ảnh hưởng đến thu hút đầu tư trực tiếp nước ngoài vào khu công nghiệp ở TPHCM".

Để có thể thu hút được các doanh nghiệp đầu tư vào các khu công nghiệp ở TPHCM ta cần tìm hiểu yếu tố môi trường đầu tư nào tác động đến khả năng thu hút đầu tư vào các khu công nghiệp, từ đó tìm giải pháp thu hút đầu tư trực tiếp từ nước ngoài vào khu công nghiệp TPHCM.

b. Mô tả vấn đề nghiên cứu

Vấn đề nghiên cứu cần được mô tả làm rõ ràng các bước sau:

Bước 1. Mục tiêu nghiên cứu

Bước 2. Câu hỏi nghiên cứu

Bước 3. Đối tượng nghiên cứu

Bước 4. Phạm vi nghiên cứu

Bước 5. Ý nghĩa và giới hạn của nghiên cứu.

Ví dụ 2. Xác định mục tiêu nghiên cứu. Đề tài: " Nghiên cứu các yếu tố ảnh hưởng đến thu hút đầu tư trực tiếp nước ngoài vào khu công nghiệp ở TPHCM".

Mục tiêu chung: Đánh giá thực trạng các khu công nghiệp về thu hút đầu tư trực tiếp nước ngoài (FDI) và xác định các yếu cố ảnh hưởng thu hút FDI trên cơ sở đó đề xuất các giải pháp nhằm thu hút FDI vào khu công nghiệp.

Mục tiệu cụ thể:

a. Đánh giá thực trạng các khu công nghiệp về thu hút FDI
b. Xác định các yếu tố ảnh hưởng thu hút FDI vào khu công nghiệp.
c. Đề xuất các giải pháp nhằm thu hút FDI vào các khu công nghiệp.

Ví dụ 3. Xác định câu hỏi nghiên cứu. Đề tài: " Nghiên cứu các yếu tố ảnh hưởng đến thu hút đầu tư trực tiếp nước ngoài vào khu công nghiệp ở TPHCM".

Câu hỏi nghiên cứu:

a. Các yếu tố môi trường đầu tư của các KCN là gì?
b. Những yếu tố nào làm thỏa mãn nhà đầu tư đang đầu tư và sẽ đầu tư tại đây?
c. Mức độ tác động của các yếu tố môi trường đầu tư đến sự thỏa mãn của nhà đầu tư như thế nào?

Ví dụ 4. Xác định đối tượng và phạm vi nghiên cứu. Đề tài: " Nghiên cứu các yếu tố ảnh hưởng đến thu hút đầu tư trực tiếp nước ngoài vào khu công nghiệp ở TPHCM".

Đối tượng nghiên cứu: Lý thuyết thực trạng của môi trường đầu tư, các yếu tố ảnh hưởng đến thu hút đầu tư trực tiếp nước ngoài vào các khu công nghiệp TPHCM.

Phạm vi nghiên cứu:

Nội dung: Đánh giá các nguồn lực cho phát triển - điều kiện tự nhiên, kinh tế, xã hội, môi trường và các KCN; môi trường đầu tư và thực trạng đầu tư vào các KCN; mức độ hài lòng và các yếu tố ảnh hưởng đến độ hài lòng của nhà đầu tư về môi trường đầu tư ở các KCN và đề xuất giải pháp thu hút đầu tư vào các KCN.

Không gian: Các khu công nghiệp trên địa bàn TPHCM.

Thời gian: Dữ liệu thu thập của 5 năm gần đây (2010 -2014) và dữ liệu do tác giả tiến hành thu thập trong năm 2014.

Tìm hiểu cơ sở lý thuyết

Cơ sở lý thuyết của nghiên cứu bao hàm các khái niệm, lý thuyết kinh tế học, các công trình nghiên cứu thực nghiệm trong và ngoài nước có liên quan:

a. Các khái niệm

Trong giai đoạn tìm hiểu về cơ sở lý thuyết, cần nêu rõ một số khái niệm trực tiếp liên quan đến vấn đề nghiên cứu. Tất cả các khái niệm được đề cập phải có nguồn gốc, trích dẫn rõ ràng để có giá trị tham khảo và phải có trong mục "tài liệu tham khảo" của đề cương nghiên cứu.

b. Các lý thuyết kinh tế học và các công trình nghiên cứu thực nghiệm

Các lý thuyết kinh tế học là những kết quả nghiên cứu đã được công nhận trong giới học thuật trên phạm vi toàn cầu, trong khi các công trình nghiên cứu thực nghiệm chỉ cho kết quả nghiên cứu riêng lẻ trong từng quốc gia, hay từng vùng, từng địa phương trong một quốc gia. Tất cả các lý thuyết được đề cập phải có nguồn gốc, trích dẫn rõ ràng để có giá trị tham khảo và phải có trong mục "tài liệu tham khảo" của đề cương nghiên cứu.

Xây dựng khung phân tích và giả thuyết nghiên cứu

a. Khung phân tích của nghiên cứu

Dựa trên lý thuyết, nghiên cứu thực nghiệm, thực tiễn địa phương, tác giả hình thành khung phân tích cho đề tài nghiên cứu của mình. Khung phân tích này cho thấy tác giả đã am tường các lý thuyết liên quan, kế thừa kết quả từ các nghiên cứu liên quan trước đây và cũng thể hiện được điểm mới trong nghiên cứu của mình. **b. Giả thuyết nghiên cứu**

Giả thuyết nghiên cứu là phát biểu về mối liên hệ giữa các biến, nhà nghiên cứu sẽ kiểm định giả thuyết này trong quá trình nghiên cứu. Dựa trên các lý thuyết kinh tế học và kế thừa kết quả từ các công trình nghiên cứu thực nghiệm, tác giả hình thành quan hệ giữa biến đôc lập và biến phụ thuộc, tương quan cùng chiều hay ngược chiều.

Viết đề cương nghiên cứu

a. Khái niệm

Đề cương nghiên cứu là một báo cáo trình bày toàn bộ các bước nghiên cứu từ vấn đề đặt ra đến tiến độ thực nghiệm nghiên cứu.

b. Nội dung của đề cương nghiên cứu

Hiện nay trên thế giới cũng như ở Việt Nam chưa có một quy định thống nhất nào về nội dung đề cương nghiên cứu. Tùy thuộc vào từng trường đại học, từng quốc gia nước khác nhau có một vài điểm giống nhau như sau:

1. Giới thiệu (mở đầu) 1.1. Xác định vấn đề nghiên cứu 1.2. Tính cấp thiết của đề tài 1.3. Mục tiêu nghiên cứu 1.4. Câu hỏi nghiên cứu 1.5. Phạm vi nghiên cứu 1.6. Ý nghĩa và giới hạn của nghiên cứu
2. Tổng quan tài liệu và cơ sở lý thuyết 2.1. Khái niệm 2.2. Lý thuyết liên quan 2.3. Các nghiên cứu trong và ngoài nước liên quan đề tài nghiên cứu
3. Khung phân tích và giả thuyết nghiên cứu 3.1. Khung phân tích của nghiên cứu 3.2. Các giả thuyết nghiên cứu
4. Phương pháp nghiên cứu (thiết kế nghiên cứu) 4.1. Phương pháp chọn mẫu 4.2. Dữ liệu thu thập 4.3. Công cụ phân tích dữ liệu
5. Kết cấu của đề tài 6. Tiến độ thực hiện 7. Tài liệu tham khảo

Bảng 1.4. Nội dung cơ bản của đề cương nghiên cứu.

Để minh họa cho đề cương nghiên cứu, chúng tôi giới thiệu chi tiết một đề cương nghiên cứu luận văn thạc sĩ của Cô Phùng Vũ Bảo Ngọc, Khoa Du lịch của trường Đại học Tài chính - Marketing (xem chương 8).

Thiết kế nghiên cứu

a. Phương pháp chọn mẫu

Phương pháp chọn mẫu bao gồm các nội dung như sau:

+) Khái niệm mẫu

+) Lý do chọn mẫu

+) Một số định nghĩa về mẫu

+) Xác định quy mô mẫu

+) Phương pháp và hình thức chọn mẫu

b. Phương pháp thu thập dữ liệu

Có hai loại dữ liệu chính để thu thập

+) Dữ liệu thứ cấp: là dữ liệu do người khác thu thập, không phải do người nghiên cứu trực tiếp thu thập.

+) Dữ liệu sơ cấp: là dữ liệu do người nghiên cứu trực tiếp thu thập.

Phân tích dữ liệu và kiểm định giả thuyết

a. Phân tích dữ liệu: Sau khi thu thập dữ liệu, ta cần tiến hành phân tích dữ liệu theo các mục tiêu và các giả thuyết nghiên cứu đưa ra.

+) Phân tích mô tả: Sử dụng thống kê mô tả để làm rõ các thuộc tính của đối tượng khảo sát.

+) Kiểm định sự khác biệt giữa hai tham số trung bình: Sử dụng các kiểm định như kiểm định t đối với mẫu đôc lập, kiểm định Chi – bình phương, kiểm định phương sai một yếu tố để xác định sự khác biệt của các tham số trung bình có ý nghĩa.

+) Kiểm định chất lượng thang đo: Sử dụng kiểm định Cronbach's Alpha để xác định chất lượng thang đo xây dựng.

+) Phân tích nhân tố khám phá: Sử dụng kiểm định KMO, Bartlett, và phương sai trích để xác định hệ thống thang đo đại diện.

+) Phân tích hồi quy tuyến tính: Sử dụng các kiểm định các hệ số hồi quy, mức độ phù hợp của mô hình, tự tương quan và phương sai phần dư để xác định các yếu tố và mức độ ảnh hưởng của các yếu tố.

b. Kiểm định giả thuyết: Sau khi có kết quả của việc phân tích dữ liệu, ta cần tiến hành kiểm định các giả thuyết đưa ra ban đầu. Có thể kết quả phân tích cho biết dữ liệu là phù hợp nhưng có thể khác với giả thuyết ban đầu. Trong trường hợp khác với giả thuyết ban đầu, ta cần điều chỉnh lại giả thuyết.

Giải thích kết quả và viết báo cáo

+) Rút ra được những phát hiện nào, kết luận nào từ kết quả?

+) Kết quả phân tích có phù hợp với lý thuyết và thực tiễn hay không?

+) Kết quả có gì mới hay không?

+) Có thể đề xuất gì hay không?

1.4. Các cấu phần cơ bản của một nghiên cứu

Cấu phần	Nội dụng	Cở sở
Ý tưởng	Nghiên cứu nhằm phát hiện điều gì và vì sao nó có ý nghĩa? +) Phát hiện khoảng trống quan trọng trong nghiên cứu trước. +) Câu hỏi nghiên cứu. +) Hình thành ý tưởng điểm mới và đóng góp của nghiên cứu	+) Hiểu biết của nhà nghiên cứu về lĩnh vực chuyên ngành. +) Tổng quan tài liệu. +) Trải nghiệm với bối cảnh cụ thể.
Thiết kế	Cần những bằng chứng gì để trả lời câu hỏi nghiên cứu và thu thập/phân tích chúng bằng cách nào? +) Cách tiếp cận và thiết kế tổng thể. +) Quy trình, phương pháp, công cụ cụ thể. +) Kế hoạch tiến hành và yêu cầu về nguồn lực, dữ liệu. +) Mức độ đáp ứng các yêu cầu của một nghiên cứu khoa học. +) Rủi ro và hạn chế của nghiên cứu.	+) Nắm rõ mục tiêu và khung nghiên cứu. +) Xác định các giả thuyết "cạnh tranh" để có kế hoạch thu thập bằng chứng, kiểm định và bảo vệ giả thuyết chính. +) Nắm rõ các phương pháp và công cụ có thể sử dụng.
Phương pháp và công cụ	Thực hiện việc thu thập và phân tích dữ liệu bằng cách nào? +) Các phương pháp thu thập dữ liệu. +) Các phương pháp và công cụ phân tích dữ liệu.	+) Nắm rõ câu hỏi và khung nghiên cứu. +) Nắm rõ quy trình và kỹ thuật của từng phương pháp.

Bảng 1.3. Các cấu phần cơ bản của một nghiên cứu.

1.5. Đề tài nghiên cứu và luận án nghiên cứu

Đề tài nghiên cứu khoa học là một nghiên cứu cụ thể có mục tiêu, nội dung, phương pháp rõ ràng nhằm tạo ra các kết quả mới, đáp ứng yêu cầu của thực tiễn sản xuất

hoặc làm luận cứ xây dựng chính sách hay cơ sở cho các nghiên cứu tiếp theo. Sau đây là một số ví dụ về đề tài nghiên cứu khoa học:

• Tình trạng nghèo đói ở Việt Nam: Thực trạng, nguyên nhân và giải pháp.
• Tìm hiểu nhu cầu và sự lựa chọn của khách hàng về sản phẩm sữa.
• Hoàn thiện tổ chức công tác kế toán tại các công ty kinh doanh dịch vụ cảng biển trên địa bàn tỉnh Bình Định. Hệ thống các đề tài nghiên cứu khoa học do Nhà nước quản lý và tài trợ thường được chia thành các cấp độ khác nhau do các cấp cơ quan quản lý khác nhau. Đề tài cấp nhà nước do Bộ Khoa học và công nghệ trực tiếp quản lý, đề tài cấp Bộ do các Bộ chuyên ngành quản lý, đề tài cấp cơ sở do các tổ chức khoa học công nghệ tự quản lý, đề tài địa phương do các Sở Khoa học và Công nghệ quản lý.

Luận án nghiên cứu là một tài liệu trình bày các kết quả nghiên cứu của tác giả nộp cho Hội đồng khoa học để ứng cử cho các chức danh sau đại học (thạc sĩ và tiến sĩ), thể hiện được kiến thức cũng như các đóng góp riêng của tác giả về chủ đề nghiên cứu.

Chương 2

ĐIỀU TRA CHỌN MẪU VÀ PHƯƠNG PHÁP CHỌN MẪU

2.1. Điều tra chọn mẫu

Mẫu và tổng thể

Tổng thể hay còn gọi là đám đông nghiên cứu là tập hợp tất cả phần tử được định nghĩa là thuộc phạm vi nghiên cứu. Tổng thể nghiên cứu là tập hợp các phần tử mà thực tế có thể nhận dạng và lấy mẫu. Mẫu nghiên cứu là một phần của tổng thể được lựa chọn để thu thập dữ liệu.

Đơn vị lấy mẫu là một hay một nhóm các phần tử để từ đó thực hiện việc lấy mẫu trong mỗi giai đoạn của quá trình chọn mẫu.

Phần tử là đơn vị mà nhà nghiên cứu cần quan sát và thu thập dữ liệu (cá nhân, hộ gia đình, tổ chức,…).

Khung mẫu là danh sách các đơn vị lấy mẫu có sẵn để phục vụ cho việc lấy mẫu.

Quy trình chọn mẫu

Bước 1. Định nghĩa tổng thể và phần tử.

Bước 2. Xác định khung lấy mẫu.

Bước 3. Xác định kích thước mẫu.

Bước 4. Xác định phương pháp chọn mẫu.

Bước 5. Tiến hành lấy mẫu theo phương pháp đã chọn.

2.2. Xác định cỡ mẫu

- Một câu hỏi luôn đặt ra với nhà nghiên cứu là cần phải điều tra bao nhiêu đơn vị mẫu để nó đại diện và có thể suy rộng cho tổng thể, để phân tích có ý nghĩa và kết quả nghiên cứu có giá trị về mặt khoa học?
- Một cách đơn giản và dễ nhất là dựa vào các nghiên cứu có cùng nội dung đã được thực hiện trước đó để lấy mẫu.
- Có thể hỏi ý kiến các chuyên gia, những người có kinh nghiệm thực hiện các dự án điều tra khảo sát.
- Có thể tính toán theo công thức tính mẫu.

- Với trường hợp cỡ mẫu lớn và không biết tổng thể.

$$n = \frac{C^2 f(1-f)}{\varepsilon^2}$$

Trong đó:

n : là cỡ mẫu (quy mô mẫu),

C : Giá trị tới hạn tương ứng với độ tin cậy lựa chọn,

Nếu biết số lượng đơn vị trong tổng thể là N, tỷ lệ tổng thể là p, sai số thì cỡ mẫu được tính theo công thức

$$n = \left[\frac{1}{N} + \frac{N-1}{N}\frac{1}{p(1-p)}\left(\frac{K}{C}\right)^2\right]^{-1}$$

Nếu tổng thể nhỏ và biết được tổng thể thì dùng công thức sau:

Với n là cỡ mẫu, N là số lượng tổng thể, e là sai số tiêu chuẩn

2.3. Các phương pháp chọn mẫu

Phương pháp chọn mẫu ngẫu nhiên đơn giản

Đây là phương pháp mà mỗi đối tượng trong tổng thể được gán một con số, sau đó các con số được lựa chọn một cách ngẫu nhiên. Thông thường các nhà nghiên cứu có thể dùng bảng ngẫu nhiên để chọn phần tử cho mẫu.

-**Ưu điểm**: Đơn giản nếu có một khung mẫu đầy đủ.

-**Nhược điểm**: Khó khả thi khi tổng thể lớn.

Ví dụ 1. Chọn 100 sinh viên trong 1000 sinh viên.

Phương pháp chọn mẫu ngẫu nhiên có hệ thống

Theo phương pháp này, toàn thể đối tượng trong tổng thể được liệt kê theo thứ tự định trước. Sau đó tùy vào quy mô mẫu và tổng thể mà quyết định khoảng cách các đối được được lựa chọn. Đây là phương pháp sử dụng phổ biến hơn phương pháp ngẫu nhiên đơn giản.

-**Ưu điểm**: không cần khung mẫu hoàn chỉnh.

-**Nhược điểm**: Mẫu sẽ bị lệch khi khung mẫu xếp theo chu kỳ và tần số bằng với bước nhảy.

Ví dụ 2. Dựa vào danh sách bầu cử tại một thành phố, ta có danh sách theo thứ tự vần của tên chủ hộ, bao gồm 240000 hộ. Ta muốn chọn ra một mẫu có 2000 hộ. Vậy khoảng cách chọn là: k= 240000/2000 = 120, có nghĩa là cứ cách 120 hộ thì ta chọn một hộ vào mẫu.

Phương pháp chọn mẫu phân tầng

Khi mẫu tương đối nhỏ, việc lựa chọn ngẫu nhiên theo hai phương pháp trên có thể dẫn tới một số đối tượng có tỷ lệ quá cao hoặc quá thấp trong mẫu. Phương pháp chọn mẫu phân tầng giúp giải quyết vấn đề này. Theo phương pháp này, các đối tượng được chia theo nhóm. Sau đó đối được được chọn ngẫu nhiên trong từng nhóm theo tỷ tệ tương ứng với tổng thể.

-**Phân tầng ngẫu nhiên theo tỷ lệ**: Số phần tử trong mỗi tầng tỷ lệ với quy mô của mỗi tầng trong tổng thể.

-**Phân tầng ngẫu nhiên không theo tỷ lệ**: Sử dụng khi độ phân tán các phần tử trong mỗi tầng khác nhau đáng kể. Số phần tử trong mỗi tầng được chọn phụ thuộc vào độ phân tán của biến quan sát trong các tầng.

Ví dụ 3. Một tòa soạn báo muốn tiến hành nghiên cứu trên một mẫu 1000 doanh nghiệp trên cả nước về sự quan tâm của họ đối với tờ báo nhằm tiếp thị việc đưa thông tin quảng cáo trên báo. Tòa soạn có thể căn cứ vào các tiêu thức: vùng địa lý (miền Bắc, miền Trung, miền Nam); hình thức sở hữu (quốc doanh, ngoài quốc doanh, công ty 100% vốn nước ngoài,…) để quyết định cơ cấu của mẫu nghiên cứu.

Phương pháp chọn mẫu theo khu vực

Trong trường hợp mà các nhóm nghiên cứu không có khả năng di chuyển quá nhiều để phỏng vấn đối tượng, họ có thể áp dụng phương pháp chọn mẫu theo khu vực. Phương pháp này không lựa chọn các đối tượng mà lựa chọn một cách ngẫu nhiên khu vực, sau đó phỏng vấn toàn bộ đối tượng trong khu vực.

Ví dụ 4. Tổng thể chung là sinh viên của một trường đại học. Khi đó ta sẽ lập danh sách các lớp chứ không lập danh sách sinh viên, sau đó chọn ra các lớp để điều tra.

Phương pháp chọn mẫu thuận tiện

Nhà nghiên cứu có thể chọn những phần tử nào mà họ có thể tiếp cận được. Nói cách khác, hình thức chọn mẫu này dựa trên sự thuận tiện hay dựa trên tính dễ tiếp cận của đối tượng khảo sát.

Ví dụ 5. Chọn mẫu n 100 doanh nghiệp bất động sản ở Thành Phố Hồ Chí Minh. Bất kỳ doanh nghiệp nào ở Thành Phố Hồ Chí Minh thuộc bất động sản và đồng ý tham gia vào mẫu đều có thể được chọn.

Phương pháp chọn mẫu phán đoán

Phương pháp chọn mẫu theo phán đoán là phương pháp mà người phỏng vấn là người tự đưa ra phán đoán về đối tượng cần chọn vào mẫu. Như vậy tính đại diện của mẫu phụ thuộc nhiều vào kinh nghiệm và sự hiểu biết của người tổ chức việc điều tra và cả người đi thu thập dữ liệu.

Ví dụ 6. Nhân viên phỏng vấn được yêu cầu đến các trung tâm thương mại chọn các phụ nữ ăn mặc sang trọng để phỏng vấn. Như vậy không có tiêu chuẩn cụ thể "thế nào là sang trọng" mà hoàn toàn dựa vào phán đoán để chọn ra người cần phỏng vấn.

Phương pháp chọn mẫu theo lớp

- Dựa vào một số thuộc tính kiểm soát xác định một số phần tử sao cho chúng đảm bảo tỷ lệ của tổng thể và các đặc trưng kiểm soát.
- Sử dụng phổ biến nhất trong thực tiễn nghiên cứu.
- Có thể dùng một hoặc nhiều thuộc tính kiểm soát như tuổi, giới tính, thu nhập, loại hình doanh nghiệp,…

Ví dụ 7. Chọn 200 hộ trong 1000 hộ ở các huyện như sau

Huyện	Số hộ trong mỗi lớp	Cỡ mẫu
A	250	50
B	150	30
C	400	80
D	200	40
Tổng	1000	200

Bảng 2.1. Thống kê lấy mẫu theo lớp.

2.4. Thiết kế bảng khảo sát

2.4.1. Những bước chính khi thiết kế bảng khảo sát

Bảng khảo sát là tập hợp các câu hỏi được trình bày theo một trình tự nhất định để người được hỏi trả lời dễ dàng và chính xác. Khi tiến hành thiết kế bảng khảo sát, người nghiên cứu cần phải trải qua 7 bước chính sau:

Bước 1. Xác định thông tin cần thu thập

Làm thế nào để xác định đầy đủ và chi tiết các thông tin cần thu thập Khi thiết kế bảng câu hỏi phải dựa vào:

-Dựa vào vấn đề nghiên cứu.

-Dựa vào nhu cầu thông tin.

-Dựa vào khung lý thuyết.

Bước 2. Xác định phương pháp phỏng vấn

Có ba phương pháp phỏng vấn chính: phỏng vấn trực diện, phỏng vấn qua điện thoại, và phỏng vấn bằng cách gửi thư/email/câu hỏi điện tử. Đối với mỗi phương pháp khác nhau người nghiên cứu sẽ xây dựng cấu trúc bảng câu hỏi khác nhau.

Bước 3. Xác định nội dung câu hỏi

Nội dung các câu hỏi thường xoay quanh việc thu thập thông tin về:

-Các sự kiện thực tế.

-Kiến thức của đối tượng được hỏi.

-Ý kiến thái độ của người đó.

-Một số dữ liệu căn bản về cá nhân đối tượng nghiên cứu để phân loại, thông tin liên lạc, và tìm kiếm các biến số liên quan.

Bước 4. Xác định hình thức câu trả lời Trả lời cho các câu hỏi đóng, gồm các dạng:

-Chọn một trong nhiều lựa chọn.

-Chọn nhiều lựa chọn.

-Xếp theo thứ tự.

Trả lời cho các câu hỏi mở

-Câu hỏi trả lời tự do.

-Câu hỏi có tính chất thăm dò.

Bước 5. Xác định cách sử dụng từ ngữ

-Nên dùng từ ngữ quen thuộc, tránh dùng tiếng lóng hoặc từ chuyên môn.

-Nên dùng từ ngữ dễ hiểu, để mọi người ở bất cứ trình độ nào cũng có thể hiểu được.

-Tránh đưa ra câu hỏi dài quá.

-Tránh đặt câu hỏi mơ hồ, không rõ ràng.

Bước 6. Xác định trình tự và hình thức bảng câu hỏi Nguyên tắc để có bảng câu hỏi đẹp:

-Mỗi phần nên được trình bày phân biệt (dùng màu giấy khác nhau).

-Đánh số các câu hỏi theo thứ tự.

-Mã hóa các phương án trả lời.

-Sử dụng dạng chữ rộng, rõ ràng.

-Đừng để các câu hỏi bị ngắt khi sang trang mới.

-Đưa ra các hướng dẫn cụ thể cho bảng hỏi nếu cần, tiếp đó là câu hỏi.

-Phân biệt giữa hướng dẫn với câu hỏi.

Bước 7. Phỏng vấn thử và hoàn thiện bảng câu hỏi

Câu hỏi đánh giá được nội dung theo mục đích đưa ra cho nó.

-Tất cả đều hiểu được câu hỏi và hiểu theo nghĩa giống nhau.

-Các hướng dẫn dễ hiểu hoặc dễ theo dõi.

-Liệu đã đưa ra hết các câu trả lời cho vấn đề chưa?

-Có thiếu câu hỏi nào một cách hệ thống/thường xuyên không?

-Kiểm tra các lỗi kỹ thuật cơ bản.

2.4.2. Những chú ý khi thiết kế từng câu hỏi

Cơ sở quan trọng khi xây dựng câu hỏi

Hai yếu tố cực kỳ quan trọng cần nắm rõ trước khi nghiên cứu:

Thứ nhất là đặc điểm của đối tượng, ví dụ trình độ học vấn, văn hóa, điều kiện kinh tế, độ tuổi,… Câu hỏi cần phù hợp với đặc điểm của đối tượng để đối tượng có thể và muốn trả lời.

Thứ hai là thông tin cần thu thập theo khung nghiên cứu. Thông tin cần thu thập là gốc để đặt câu hỏi. Tuy nhiên, câu hỏi không nhất thiết hỏi thẳng vào thông tin cần mà phải hỏi những thông tin mà đối tượng có thể trả lời.

Các loại câu hỏi

a. Phân theo hình thức, có câu hỏi đóng, câu hỏi có nhiều lựa chọn và câu hỏi mở

Câu hỏi đóng đơn giản: là dạng câu hỏi chỉ có hai thái cực trả lời như "Có"/ "Không", "Đúng"/ "Sai",…

Câu hỏi có lựa chọn định sẵn và đối tượng có thể chọn nhiều phương án phù hợp: Đây là một dạng khác của câu hỏi đóng đơn giản khi bản thân mỗi phương án là một câu hỏi đóng.

Câu hỏi có lựa chọn định sẵn nhưng chỉ chọn một phương án.

Câu hỏi mở: Dạng câu hỏi này không có các phương án để lựa chọn mà đối tượng có thể điền câu trả lời theo ý của mình. Câu hỏi mở được sử dụng hạn chế trong khảo sát định lượng vì sẽ mất công mã hóa.

b. Phân theo nội dung, câu hỏi có thể chia làm ba loại

- Câu hỏi về thông tin khách quan.
- Câu hỏi về hành vi hoặc trải nghiệm cụ thể.
- Câu hỏi về cảm nhận, thái độ và đánh giá của đối tượng.

2.4.3. Những chú ý khi thiết kế tổng thể bảng câu hỏi

Thiết kế tổng thể bảng câu hỏi cũng là một công đoạn quan trọng để đảm bảo đối tượng muốn trả lời bảng câu hỏi. Có một số kinh nghiệm khi thiết kế bảng câu hỏi như sau

Hình thức

Bảng câu hỏi cần được trình bày cẩn thận, dễ nhìn và nhất quán. Việc thiết kế cũng đảm bảo thuận lợi cho đối tượng lựa chọn và điền câu trả lời.

Giới thiệu

Bảng câu hỏi nên có phần giới thiệu hoặc thư giới thiệu đính kèm. Phần giới thiệu cần nêu mục đích cuộc khảo sát (không nhất thiết phải quá cụ thể - nên dừng ở mức mà đối tượng quan tâm). Phần này cũng nên khẳng định việc bảo mật danh tính người trả lời và cung cấp địa chỉ liên hệ của nhóm nghiên cứu.

Các câu hỏi cơ bản

Có thể phân chia câu hỏi theo các phần để đối tượng dễ trả lời. Nên bắt đầu bằng những phần dễ trả lời, ít nhạy cảm.

Trong một số trường hợp đối tượng trả lời có thể bỏ qua một số câu hỏi. Khi đó, việc hướng dẫn chuyển câu hỏi cần được ghi rõ ràng (ví dụ: Nếu trả lời "Không", chuyển sang câu 10).

Khi có các câu hỏi nhạy cảm, nên đan xen với những câu hỏi ít nhạy cảm hơn.

Ưu tiên các câu hỏi về thông tin khách hàng, sau đó đến câu hỏi về trải nghiệm và hành vi. Các câu hỏi về cảm nhận và đánh giá có ưu tiên thấp hơn, trừ khi chính cảm nhận và đánh giá của đối tượng là mục tiêu cần nghiên cứu.

Các câu hỏi theo nhóm

Các câu hỏi phân nhóm thường là đặc điểm của đối tượng trả lời (Cá nhân, hộ gia đình, doanh nghiệp,…). Các thông tin này dùng để phân nhóm, so sánh nhóm và để kiểm soát khi sử dụng các mô hình kiểm định thống kê.

Độ dài bảng câu hỏi

Độ dài bảng câu hỏi phụ thuộc vào hai yếu tố cơ bản: thông tin cần thu thập và nguồn lực của đề tài. Một bảng câu hỏi quá dài thường khó thuyết phục các đối tượng trả lời. Ngược lại, một bảng câu hỏi quá ngắn có thể không thu thập đủ thông tin cần thiết. Khi không có lợi ích đi kèm (ví dụ : quà tặng), một đối tượng có thể chỉ sẵn sàng dành 20 – 25 phút để trả lời bảng câu hỏi.

2.4.4. Xây dựng và thực hiện quy trình khảo sát

Sau khi xác định mẫu khảo sát và xây dựng bảng khảo sát, bước tiếp theo là xác định quy trình khảo sát để đảm bảo thu thập được thông tin với độ tin cậy cao.

Đối với khảo sát qua thư, quy trình như sau:

Kiểm tra lại địa chỉ liên hệ của đối tượng.

Tiến hành gởi thư tới các đối tượng.

Gọi điện thông báo trước.

Gửi thư cảm ơn.

Đối với khảo sát bằng phương pháp phỏng vấn, quy trình như sau:

Tập huấn cho các bộ phỏng vấn: Phỏng vấn trực tiếp đòi hỏi có nhiều người cùng tham gia cùng phỏng vấn. Các cán bộ phỏng vấn cần thực hiện đúng quy trình phỏng vấn một cách nhất quán.

Gọi điện liên hệ và hẹn thời gian phỏng vấn.

Tiến hành phỏng vấn: Nhóm nghiên cứu cần đảm bảo địa điểm phỏng vấn thuận lợi cho việc trả lời một cách khách quan.

Giám sát và đảm bảo chất lượng.

2.4.5. Quy trình chuẩn bị số liệu

Nhập liệu

Công đoạn này cần thiết đối với phương pháp thu thập qua thư hoặc phỏng vấn trực tiếp bằng phiếu giấy. Cần lưu ý:

-Mỗi dòng được dành cho một quan sát (thường là một phiếu).

-Mỗi cột là một trường dữ liệu.

-Mỗi phiếu câu hỏi gán một mã.

-Nhập dữ liệu theo trình tự câu hỏi và trung thành với giá trị trong bảng câu hỏi.

Không tiến hành điều chỉnh khi nhập số liệu trừ khi nhận rõ sai sót khi nhập số liệu.

-Nhập phiếu hai lần độc lập.

-File dữ liệu có thể được kiểm tra bằng các lệnh tần suất đơn giản. Nếu có các giá trị nằm ngoài khoảng cho phép hoặc đáng ngờ thì nhóm nghiên cứu có thể đối chiếu lại với phiếu câu hỏi.

Kiểm định các thước đo

Các biến số về thái độ, hành vi, hay cảm nhận thường được đo lường bằng một số câu hỏi hoặc mệnh đề. Kể cả khi những thước đo được kiểm định cẩn thận ở những nghiên cứu trước đó, đối với mỗi cuộc khảo sát, những thước đo này vẫn cần được kiểm tra về độ tin cậy.

- Phân tích nhân tố (factor analysis) : Phân tích nhân tố chính là việc kiểm tra xem các mệnh đề/câu hỏi có thực sự nhóm lại với nhau thành thước đo như trong lý thuyết hay không. Với các khảo sát khác nhau, có thể một số mệnh đề không vào cùng nhóm với các mệnh đề khác. Khi đó nhóm nghiên cứu cần tiếp tục kiểm tra độ tin cậy để ra quyết định.
- Phân tích độ tin cậy (Reliability analysis) : Phân tích độ tin cậy là xem các mệnh đề có thực sự "thống nhất" với nhau để cùng đo lường biến số cần đo hay không. Chỉ số đo lường sự thống nhất này là Cronbach's alpha. Chỉ số này tốt là từ 0,7 trở lên và tối thiểu cần đạt là 0,63 (D'Vellis, 1990).

Các phần mềm thống kê có thể giúp thực hiện hai phép phân tích này khá nhanh chóng và dễ dàng. Bạn đọc có thể tìm hiểu kỹ thuật phân tích nhân tố ở các sách vở viết về thống kê toán.

2.4.6. Câu hỏi thảo luận

Câu hỏi 1. Nghiên cứu khảo sát là gì? Khi nào ta cần áp dụng nghiên cứu khảo sát? Nêu một ví dụ áp dụng phương pháp khảo sát?

Câu hỏi 2. Khi thiết kế khảo sát cần lưu ý những vấn đề gì? Nêu những hạn chế của phương pháp khảo sát?

Câu hỏi 3. Mẫu khảo sát cần đảm bảo yêu cầu gì? Yếu tố nào ảnh hưởng tới tính đại diện của mẫu.

Câu hỏi 4. Hãy nêu các phương pháp chọn mẫu? Phương pháp khảo sát nên chọn phương pháp chọn mẫu nào?

Câu hỏi 5. Trình bày phương pháp lấy mẫu theo xác suất? cho ví dụ minh họa.

Câu hỏi 6. Những loại nghiên cứu nào thì sử dụng phương pháp khảo sát?

Chương 3

KIỂM ĐỊNH GIẢ THUYẾT NGHIÊN CỨU

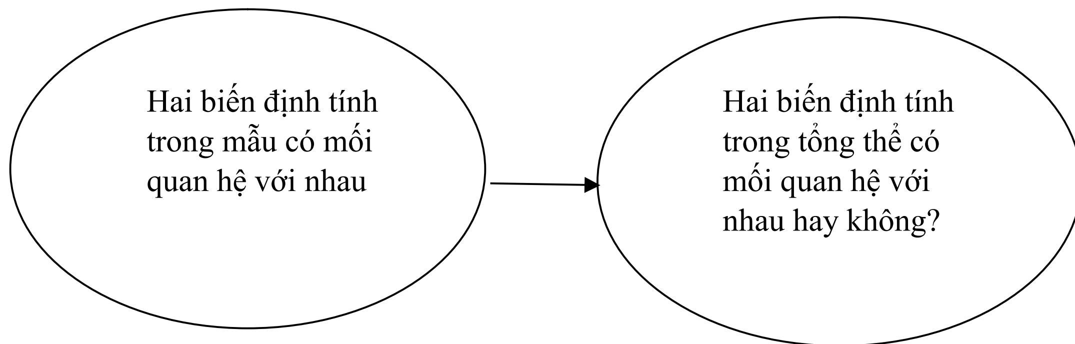
3.1. Một số khái niệm

Giả thuyết thống kê là những phát biểu về các tham số, quy luật phân phối, hoặc tính độc lập của các đại lượng ngẫu nhiên. Việc tìm ra kết luận để bác bỏ hay chấp nhận một giả thuyết được gọi là kiểm định thống kê.

Mục đích của kiểm định giả thuyết là để rút ra kết luận về các tham số của tổng thể từ kết quả quan sát của mẫu được chọn ngẫu nhiên.

Phương pháp kiểm định: Thứ nhất là đưa ra cặp giả thiết đối gồm giả thiết không (Null Hypothesis) H_0 và giả thiết đối H_1. Thứ hai là chọn một kiểm định thống kê để đánh giá giả thiết không H_0.

3.2. Kiểm định Chi-square



Giả sử hai biến định tính trong mẫu có mối quan hệ với nhau, ta muốn kiểm định xem hai biến định tính này có mối quan hệ với nhau trong tổng thể hay không. Khi đó, ta thực hiện kiểm định Chi-square.

Cặp giả thiết:

H_0: Hai biến độc lập với nhau

H_1: Hai biến có liên hệ với nhau

Tiêu chuẩn kiểm định:

- Nếu sig. < 0.05 thì ta bác bỏ giả thiết H_0 với độ tin cậy 95%. Ngược lại, ta chấp nhận giả thiết H_0 với độ tin cậy 95%.

Ví dụ minh họa: (file ***Datathuchanh.sav***)

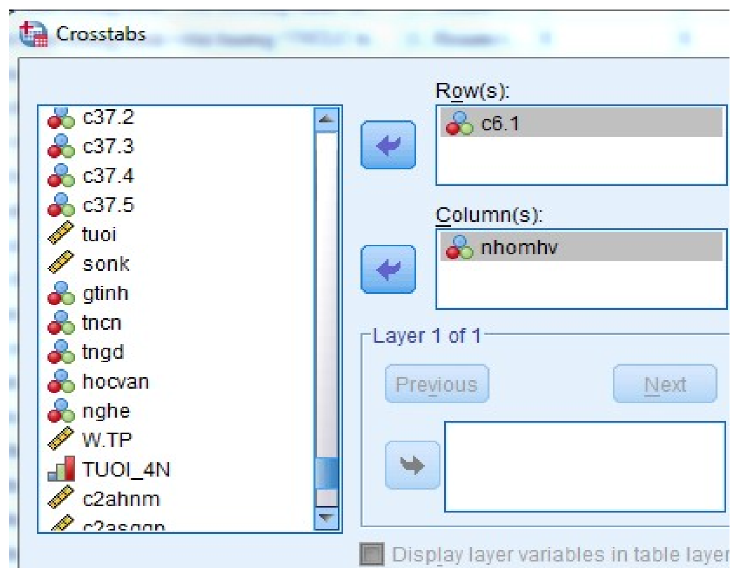
Với mẫu gồm 500 quan sát, dựa vào bảng kết hợp hai biến định tính dưới đây (Cách thức đọc báo nói chung và trình độ học vấn), ta thấy rằng nhóm người có trình độ học vấn cao có xu hướng đọc báo theo cách "Xem lướt qua các đề mục, đọc các mục ưa thích trước". Nhóm người có trình độ học vấn cấp 1, cấp 2 thì "Chỉ đọc các trang mục ưa thích, ít đọc các trang khác". Như vậy, dựa vào mẫu này, rõ ràng là có mối quan hệ giữa cách thức đọc

báo và trình độ học vấn của người đọc. Tuy nhiên, mục tiêu nghiên cứu của chúng ta là tổng thể, chứ không phải mẫu, do đó, chúng ta cần thực hiện kiểm định. Trong file ***Datathuchanh.sav,*** ta tiến hành kiểm định biến c61 (Cách đọc các tờ báo nói chung) và biến nhomhv (Trình độ học vấn) để xem thử hai biến này có mối liên hệ trong tổng thể không?

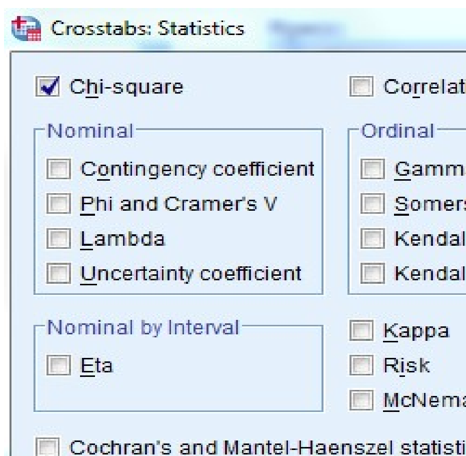
		Nhom hoc van					
		Cap 1-2		Cap 3		CD-SVDH	
		Count	Column N %	Count	Column N %	Count	Column N
Cach doc cac to bao noi chung	Doc theo thu tu tu trang dau den trang cuoi	18	30.0%	77	35.8%	18	19.
	Xem luot qua cac de muc, doc cac muc ua thich truoc	23	38.3%	95	44.2%	57	62.
	Chi doc cac trang muc ua thich, it doc cac trang khac	9	15.0%	20	9.3%	9	9.

Thực hiện với SPSS:

B1: Từ menu, chọn **Analyze > Descriptive Statistics > Crosstabs…** Đưa biến c6.1 vào ô **Row(s):** và đưa biến nhomhv vào ô **Column(s):**



B2: Nhấn nút **Statistics** để chọn kiểm định Chi-square. Nhấn **Continue > OK.**



Chi-Square Tests

	Value	df	Asy (2
Pearson Chi-Square	19.336[a]	9	
Likelihood Ratio	19.215	9	
Linear-by-Linear Association	.261	1	
N of Valid Cases	495		

Ta được kết quả như trên. Ta thấy rằng, ứng với đại lượng thống kê Pearson Chi-Square thì P-value = 0.022 < 0.05 nên ta bác bỏ giả thuyết H_0 với độ tin cậy 95%. Trong SPSS, P-value chính là Sig. (Observed Significance Level).

Chú ý rằng kiểm định Chi-square dùng để kiểm định mối liên hệ giữa hai biến định danh-định danh hoặc định danh-thứ bậc. Trong trường hợp hai biến thu thập từ thang đo thứ bậc thì thay vì dùng kiểm định Chi-square, chúng ta có thể dùng **Tau của Kendall, Gamma của Goodman và Kruskal.**

3.3. Kiểm định giả thuyết về giá trị trung bình của tổng thể

Trong thống kê, có các phép kiểm định về giá trị trung bình của tổng thể phổ biến sau:

TT	Mục đích kiểm định	SPSS
1	So sánh giá trị trung bình của một tổng thể với một giá trị cụ thể	Analyze > Compare Means > One-Sample T-Test
2	So sánh hai giá trị trung bình của hai tổng thể riêng biệt	Analyze > Compare Means > Independent-Samples T-Test

3	So sánh hai giá trị trung bình của hai tổng thể riêng biệt, nhưng mỗi quan sát trong tổng thể này có sự tương đồng theo cặp với một quan sát trong tổng thể kia.	Analyze > Compare Means > Paired-Samples T-Test
4	So sánh các giá trị trung bình của nhiều hơn hai tổng thể riêng biệt	Analyze > Compare Means > One-way ANOVA

So sánh giá trị trung bình của một tổng thể với một giá trị cụ thể

Tình huống: Có người cho rằng tuổi trung bình của độc giả báo Sài Gòn Tiếp thị (SGTT) là 30 tuổi. Vỡi dữ liệu có được trên file ***Datathuchanh.sav.*** Bạn làm thế nào để kiểm định giả thuyết này?

Cách thức kiểm định:

B1: Đặt giả thuyết H_0: Tuổi trung bình của độc giả báo Sài Gòn Tiếp thị = 30

B2: Dùng lệnh **Transform → Count values within cases** để chuyển các biến **Category** từ **c2a.1 đến ca2.9** thành biến **Dichotomy** tên là ***docSGTT*** với biểu hiện 1 nếu là người có đọc báo SGTT và biểu hiện 0 nếu là người không đọc báo SGTT mà đọc các báo khác. (Xem hướng dẫn chi tiết trong file **Chủ đề 2 HD Category thanh Dichotomy.doc**)

B3: Dùng lệnh **Seclect Case** để lọc ra những trường hợp mà ***docSGTT*** nhận giá trị 1 để các lệnh thống kê sau đó của SPSS chỉ thực hiện trên những trường hợp này.

Data thuc hanh.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Define Variable Properties...
Set Measurement Level for Unknown...
Copy Data Properties...
New Custom Attribute...
Define Dates...
Define Multiple Response Sets...
Validation
Identify Duplicate Cases...
Identify Unusual Cases...
Sort Cases...
Sort Variables...
Transpose...
Merge Files
Restructure...
Aggregate...
Orthogonal Design
Copy Dataset
Split File...
Select Cases...

	N	Label	Values	
146	sonk	haân khaåu trong hoä	None	N
147	gtinh	nh	{1, Nam}...	N
148	tncn	ù nhaân TB thaùng	{1, Khoâng...	N
149	tngd	ä TB thaùng	{1, Döôùi 2t...	6
150	hocva	do hoc vanvaán	{1, Caáp 1}...	N
151	nghe	nghiep	{1, Cong ch...	N
152	W.TP		None	N
153	TUOI	oi	{1, 18-25}...	N
154	nhom	hoc van	{1, Cap 1-2}...	N
155	c2ahn	c HNM	None	N
156	c2asg	c SGGP	None	N
157	c2ald	c LD	None	N
158	c2anl	c NLD	None	N
159	c2atp	c TP	None	N
160	c2atn	c TN	None	N
161	c2att	c TT	None	N
162	DocT	oitre	{.00, Khong...	N
163	filter_	TT = 1 (FILTER)	{0, Not Sele...	N
164	docS	GTT	{0, Khong d...	N
165				

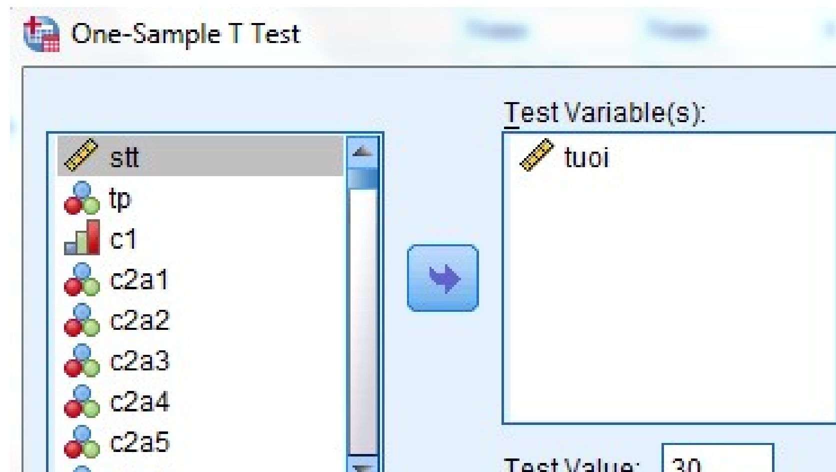
Select Cases

tuoi
sonk
gtinh
tncn
tngd
hocvan
nghe
W.TP
TUOI_4N
nhomhv
c2ahnm
c2asggp
c2ald
c2anld
c2atp
c2atn
c2att
DocTT
filter_$

Select
All cases
If condition is satisfied
If... docSGTT = 1
Random sample of cases
Sample...
Based on time or case range
Range...
Use filter variable:

Output
Filter out unselected cases
Copy selected cases to a new datas
Dataset name:
Delete unselected cases

B4: Trên menu, chọn **Analyze > Compare Means > One-Sample T-Test.** Trên hộp thoại **One-Sample T-Test**, đưa biến tuoi vào khung **Test variable**, khai báo **Test Value** là 30. Nhấn nút **Options…** để chọn độ tin cậy.



Nhấn nút **Continue > OK**. Ta nhận được bảng kết quả sau:

T-Test

```
[DataSet1] D:\Thanh Phuong\SPSS SV\Data thuc hanh.sav
```

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Tuoi	159	32.79	10.328	.819

One-Sample Test

	Test Value = 30				
					99% Confide

Theo mẫu này (N = 159), tuổi trung bình của các độc giả SGTT là 32.79 tuổi. Mức ý nghĩa của kiểm định Sig. = 0.001 < 0.01, do đó ta bác bỏ giả thuyết H_0 với độ tin cậy 99%.

2.786 = Mean Difference = Mean của mẫu (32.79) – Mean giả thuyết (30)

0.65 < Mean Difference < 4.92 với độ tin cậy 99%.

Do đó, ta có thể khẳng định tuổi trung bình của những người thường xuyên đọc báo Sài Gòn Tiếp thị là trên 30 tuổi.

So sánh hai giá trị trung bình của hai tổng thể riêng biệt-Trường hợp mẫu độc lập

Tình huống: Cần so sánh về số nhân khẩu trung bình của hộ gia đình giữa hai thành phố Hà Nội và Thành phố Hồ Chí Minh. Với dữ liệu có được trên file ***Datathuchanh.sav.*** Bạn làm thế nào để thực hiện điều này?

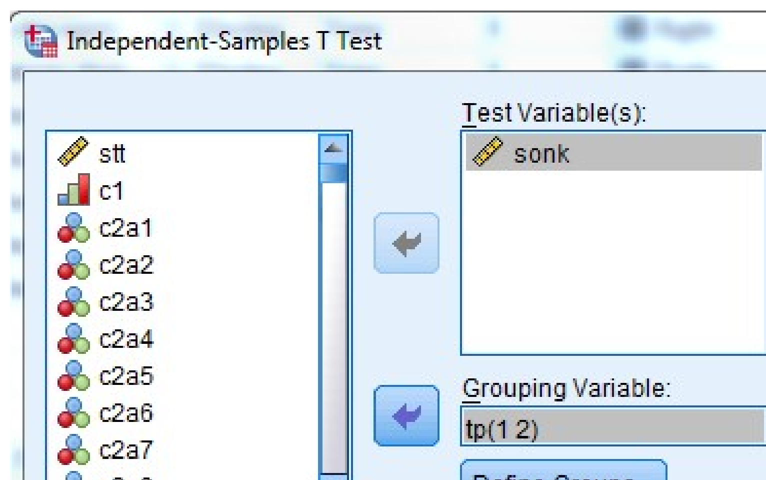
Phân tích tình huống: Trong tình huống này, bạn có một biến định lượng là *sonk* và một biến định tính là *tp*. Từ thông tin của hai mẫu độc lập, bạn sẽ sử dụng kiểm định sự bằng nhau về giá trị trung bình của hai tổng thể có phân phối chuẩn và phương sai bằng nhau. Từ đây về sau, ta gọi tắt là kiểm định trung bình hai mẫu độc lập (Independent-Samples T-Test). Nếu kiểm định giả thuyết về giá trị trung bình của hai mẫu độc lập vi phạm giả định là hai mẫu được lấy ngẫu nhiên từ hai tổng thể có phân phối chuẩn thì chúng ta phải thực hiện một phương pháp kiểm định khác tên là Mann-Whitney.

Trước khi thực hiện kiểm định trung bình, ta cần thực hiện **kiểm định sự bằng nhau của hai phương sai tổng thể**. Kiểm định này có tên là **Levene Test**.

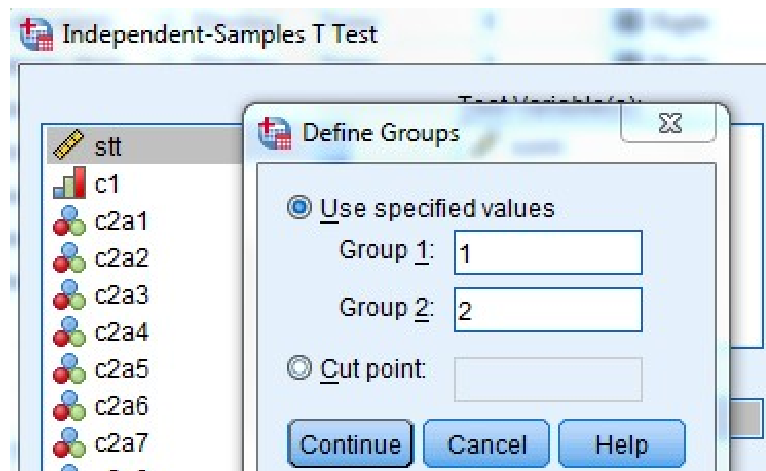
Cách thức kiểm định:

B1: Đặt giả thuyết H_0: Số nhân khẩu trung bình của hộ gia đình tại hai thành phố là bằng nhau.

B2: Trên menu, chọn Analyze > Compare Means > Independent-Samples T-Test. Trong hộp thoại, chọn biến định lượng *sonk* vào mục Test Variable(s):. Chọn biến định tính *tp* vào mục Grouping Variable:



Sau khi chọn biến định tính để phân nhóm, nhấn nút Define Groups… để chỉ định hai nhóm cần so sánh với nhau. Trong hộp thoại Define Groups…, gõ mã số của hai nhóm cần so sánh.



B3: Nhấn nút Continue > OK. Ta nhận được bảng kết quả sau:

T-Test

[DataSet1] D:\Thanh Phuong\SPSS SV\Data thuc hanh.sav

Group Statistics

	Thanh pho	N	Mean	Std. Deviation	Std. Error Mean
So nhan khau	Ha Noi	250	4.29	1.667	.105
	TPHCM	250	5.33	2.715	.172

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means				
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference

Dựa vào kết quả kiểm định sự bằng nhau của hai phương sai **(Levene's Test for Equality of Variances**), ta sẽ xem xét kết quả của kiểm định t, cụ thể như sau:

+ Nếu giá trị Sig. trong kiểm định Levene < 0.05 thì phương sai giữa hai thành phố khác nhau, ta sẽ sử dụng kết quả kiểm định t ở phần Equal Variances not Assumed.

+ Ngược lại, nếu giá trị Sig. trong kiểm định Levene >= 0.05 thì phương sai giữa hai thành phố không khác nhau, ta sẽ sử dụng kết quả kiểm định t ở phần Equal Variances Assumed.

Với kết quả kiểm định về sự bằng nhau của hai phương sai ở bảng trên, ta bác bỏ giả thuyết H_0 về sự bằng nhau của hai phương sai. Do đó, chúng ta sẽ sử dụng kết quả ở phần Equal Variances not Assumed cho kiểm định t, cụ thể như sau:

+ Nếu giá trị Sig. trong kiểm định t < 0.05 thì ta kết luận có sự khác nhau về số nhân khẩu trung bình của hộ gia đình giữa hai thành phố.

+ Nếu giá trị Sig. trong kiểm định t >= 0.05 thì ta kết luận chưa có sự khác nhau về số nhân khẩu trung bình của hộ gia đình giữa hai thành phố.

Trong ví dụ này, vì Sig. < 0.05 nên ta bác bỏ giả thuyết H_0 về sự bằng nhau của hai trung bình và kết luận số nhân khẩu trung bình trong hộ gia đình ở Thành phố Hồ Chí Minh lớn hơn so với ở Hà Nội một cách có ý nghĩa thống kê (5.33 > 4.29).

So sánh hai giá trị trung bình của hai tổng thể riêng biệt, nhưng mỗi quan sát trong tổng thể này có sự tương đồng theo cặp với một quan sát trong tổng thể kia (Paired-Samples T-Test)

Dau phong.sav [DataSet2] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct

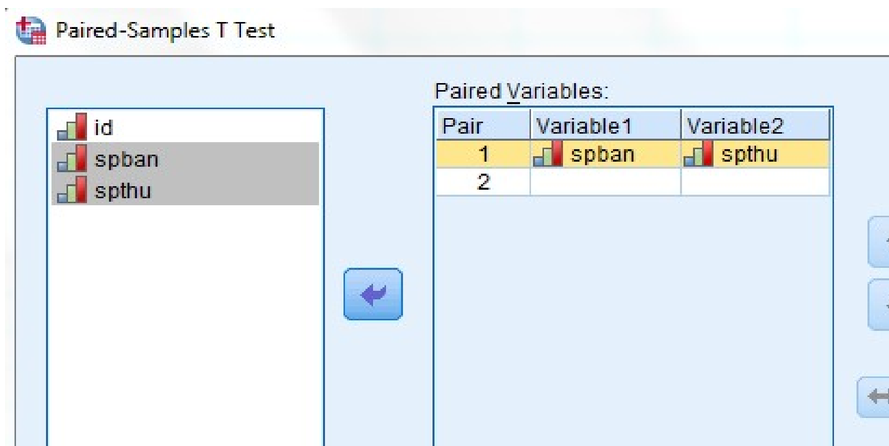
	id	spban	spthu
1	1	7	
2	2	8	
3	3	6	
4	4	8	
5	5	7	
6	6	7	
7	7	7	
8	8	6	
9	9	8	
10	10	6	
11	11	7	
12	12	7	
13	13	8	
14	14	9	
15	15	7	

Tình huống: Một công ty chế biến thực phẩm cần kiểm định điểm đánh giá trung bình của người tiêu dùng về đậu phộng khi chưa được cải tiến và khi đã được cải tiến. Vỡi dữ liệu có được trên file ***Dauphong.sav.*** Bạn làm thế nào để kiểm định giả thuyết này?

Cách thức kiểm định:

B1: Đặt giả thuyết H_0: Không có sự khác nhau về điểm trung bình tổng thể

B2: Trên menu, chọn Analyze > Compare Means > Paired-Sample T-Test. Chọn hai biến *spban* và *spthu* vào mục **Paired Variables:**



B3: Nhấn nút Continue > OK. Ta nhận được bảng kết quả sau:

T-Test

[DataSet2] D:\Thanh Phuong\SPSS SV\Dau phong.sav

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	san pham dang ban	7.20	20	.834	.186
	san pham cai tien	7.80	20	1.399	.313

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	san pham dang ban & san pham cai tien	20	.533	.016

Paired Samples Test

	Paired Differences

Trong bảng cuối cùng, vì Sig. = 0.36 < 0.05 nên có thể kết luận rằng có sự chênh lệch có ý nghĩa thống kê về sự đánh giá của người tiêu dùng đối với sản phẩm đậu phộng trước và sau khi cải tiến. Cụ thể là sản phẩm sau khi cải tiến được đánh giá cao hơn sản phẩm trước cải tiến, trung bình khoảng 0.6 điểm.

3.4. Phân tích phương sai một yếu tố (Analysis of Variance-ANOVA)

Phân tích phương sai là sự mở rộng của kiểm định t vì phương pháp này giúp ta so sánh giá trị trung bình của 3 nhóm trở lên. Sau đây là một số giả định đối với phân tích phương sai một yếu tố:

+ Các nhóm so sánh phải độc lập và được chọn một cách ngẫu nhiên.

+ Các nhóm so sánh phải có phân phối chuẩn hoặc cỡ mẫu phải đủ lớn để được xem như tiệm cận phân phối chuẩn.

+ Phương sai của các nhóm so sánh phải đồng nhất.

Nếu giả định tổng thể có phân phối chuẩn với phương sai bằng nhau không được đáp ứng thì **kiểm định phi tham số Kruskal-Wallis** sẽ là một giải pháp thay thế hữu hiệu cho ANOVA.

3.5. Phân tích tình huống và cách thức thực hiện trên SPSS

Tình huống: Có sự khác biệt hay không về mức độ đánh giá tầm quan trọng của yếu tố "có tự do cá nhân" đối với cuộc sống của một con người giữa những người có trình độ học vấn khác nhau. Biết rằng

+ Biến định lượng *c36.6* được xác định bởi thang đo định lượng 7 mức độ (1: không quan trọng, 7: rất quan trọng)

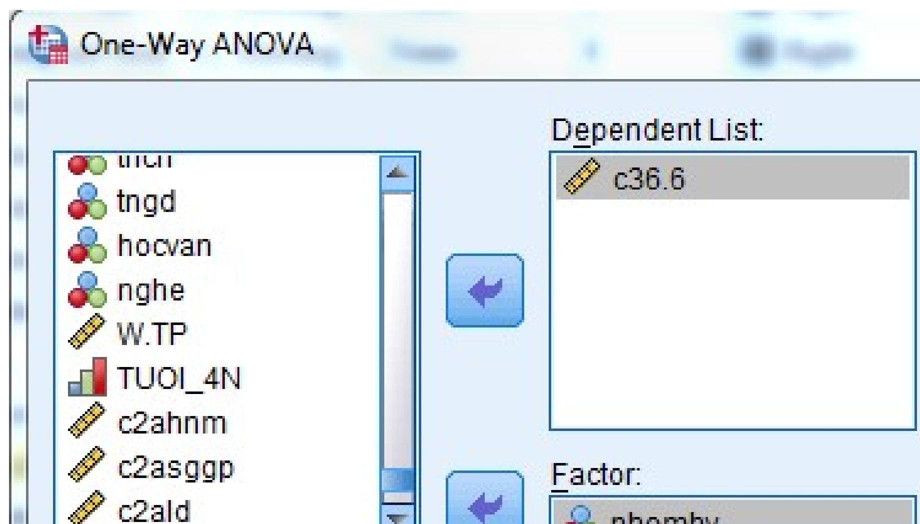
+ Biến định tính *nhomhv* (1: cấp 1-2, 2: cấp 3-THCN, 3: CĐ-SVĐH, 4: Tốt nghiệp ĐH)

Với dữ liệu có được trên file ***Datathuchanh.sav.*** Bạn làm thế nào để kiểm định giả thuyết này?

Cách thức kiểm định:

B1: Đặt giả thuyết H_0: Không có sự khác biệt về mức độ đánh giá tầm quan trọng của yếu tố "có tự do cá nhân" giữa các nhóm trình độ học vấn

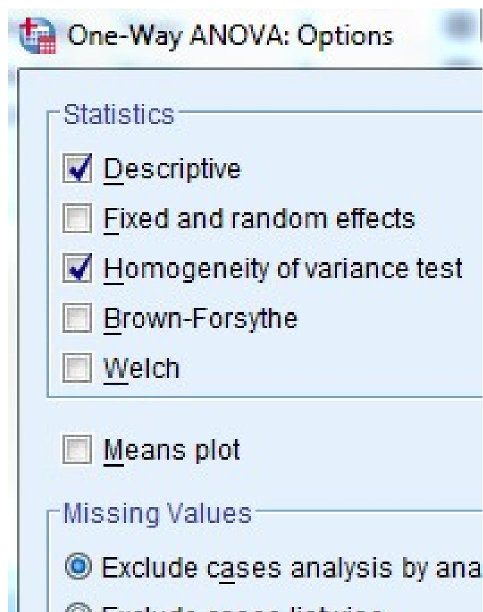
B2: Trên menu, chọn Analyze > Compare Means > One-way ANOVA. Đưa biến định lượng vào mục **Dependent List** và đưa biến phân loại vào khung **Factor**.



B3: Chọn nút **Options…**

Descriptive: Tính các đại lượng thống kê mô tả chi tiết cho từng nhóm được phân tách để so sánh.

Homogeneity-of-variance: Kiểm định sự bằng nhau của các phương sai nhóm.



Nhấn nút **Continue > OK.** Ta có các bảng kết quả sau:

Oneway

[DataSet1] D:\Thanh Phuong\SPSS SV\Data thuc hanh.sav

Descriptives

Co tu do ca nhan

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
					Lower Bound	Upper Bound
Cap 1-2	60	5.43	1.294	.167	5.10	5.77
Cap 3	215	5.18	1.391	.095	4.99	5.37

Test of Homogeneity of Variances

Co tu do ca nhan

Levene Statistic	df1	df2

ANOVA

Co tu do ca nhan

	Sum of Squares	df	Mean Square	F
Between Groups	13.488	3	4.496	2.2
Within Groups	963.413	491	1.962	

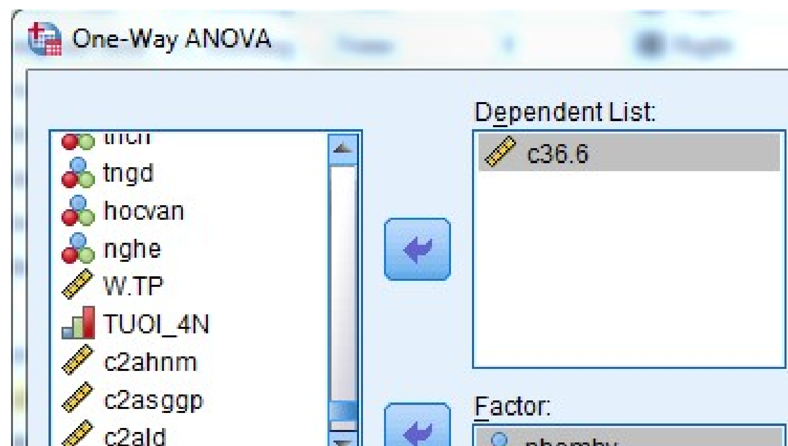
Bảng thứ nhất cho thông tin về các đại lượng thống kê mô tả cho từng nhóm và cho toàn bộ mẫu nghiên cứu.

Bảng thứ hai cho biết kiểm định phương sai. Với Sig. = 0.258 có thể nói phương sai của sự đánh giá tầm quan trọng của yếu tố "có sự tự do cá nhân" giữa 4 nhóm học vấn không khác nhau một cách có ý nghĩa thống kê. Như vậy, kết quả phân tích ANOVA ở bảng thứ ba được sử dụng tốt. Ở bảng thứ ba, với Sig. = 0.077, nếu chọn độ tin cậy của kiểm định là 90 % thì có thể nói có sự khác biệt có ý nghĩa thống kê về sự đánh giá tầm quan trọng của yếu tố "có sự tự do cá nhân" giữa 4 nhóm học vấn **(0.077 < 0.1).** Nếu chọn độ tin cậy của kiểm định là 95 % thì có thể nói không có sự khác biệt có ý nghĩa thống kê về sự đánh giá tầm quan trọng của yếu tố "có sự tự do cá nhân" giữa 4 nhóm học vấn **(0.077 > 0.05).**

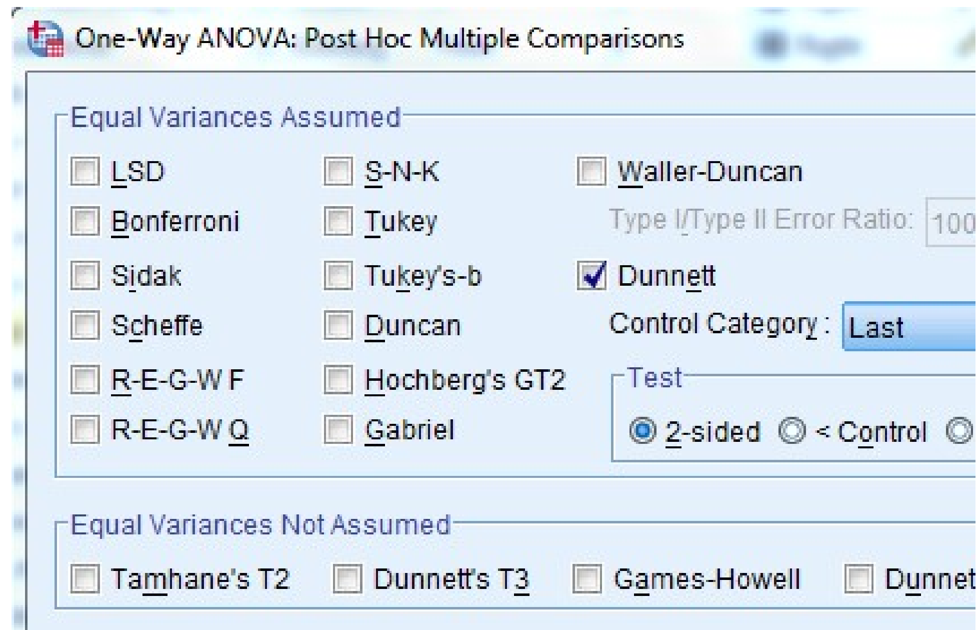
3.5.1. Xác định chỗ khác biệt (Phân tích sâu ANOVA)

Việc tiếp theo là chúng ta cần tìm xem sự đánh giá này là khác biệt giữa các nhóm học vấn nào. Có hai phương pháp để xá định sự khác biệt trong các giá trị trung bình nhóm xảy ra ở đâu, đó là kiểm định "trước" và kiểm định "sau".

\+ Kiểm định "trước" có thể định nghĩa như là kiểm định các giả định về sự khác nhau của các trung bình nhóm trước khi thực hiện phân tích ANOVA. Lúc này ta thực hiện việc xác định những chênh lệch cụ thể giữa các trung bình nhóm theo phán đoán chủ quan của ta. Kiểm định này được thực hiện trong hộp thoại Contrasts với tên gọi là kiểm định Priori Contrasts.



+ Kiểm định “sau” có thể định nghĩa như là kiểm định các giả định về sự khác nhau của các trung bình nhóm sau khi thực hiện phân tích ANOVA. Kiểm định này được thực hiện trong hộp thoại Post Hoc. Sau khi nhấn Post Hoc, ta có hộp thoại sau:



Trong ví dụ này, ta thử chọn kiểm định Dunnett với lựa chọn mặc định là nhóm cuối cùng. Sau đó nhấn nút Continue > OK. Ta được bảng kết quả sau:

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Co tu do ca nhan
Dunnett t (2-sided)[a]

(I) Nhom hoc van	(J) Nhom hoc van	Mean Difference (I-J)	Std. Error	Sig.	90% Co Lower Bou
Cap 1-2	Tot nghiep DH	.480*	.219	.077	.
Cap 3	Tot nghiep DH	.228	.156	.337	-.
CD-SVDH	Tot nghiep DH	.409*	.192	.088	

Dựa vào bảng kết quả trên, ta thấy có sự khác biệt có ý nghĩa thống kê giữa nhóm có trình độ học vấn cấp 1-2 và nhóm đã tốt nghiệp Đại học vì Sig. = 0.077 < 0.1.

3.5.2. Phân tích phương sai hai yếu tố (Two-way ANOVA)

(file Phan tich phuong sai 2 yeu to.sav)

Sau đây là một số giả định đối với phân tích phương sai hai yếu tố:

+ Tổng thể có phân phối chuẩn.

+ Phương sai của các nhóm so sánh phải đồng nhất.

Tình huống: Xét ảnh hưởng của mức độ yêu thích ngành học và thời gian tự học đến kết quả học tập của sinh viên.

Cách thức kiểm định:

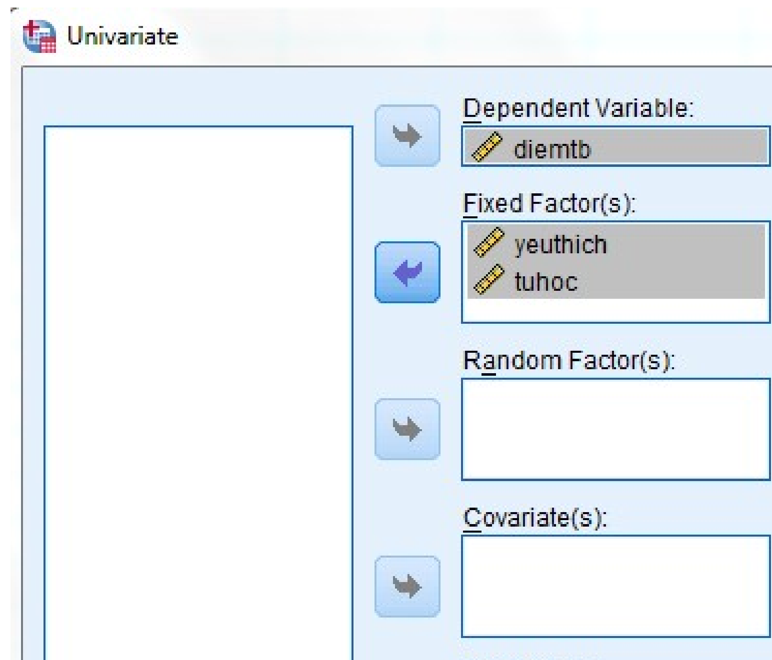
B1: Đặt các giả thuyết H_0 cho tình huống này như sau:

+ Điểm trung bình (ĐTB) của sinh viên có thời gian tự học khác nhau đều bằng nhau.

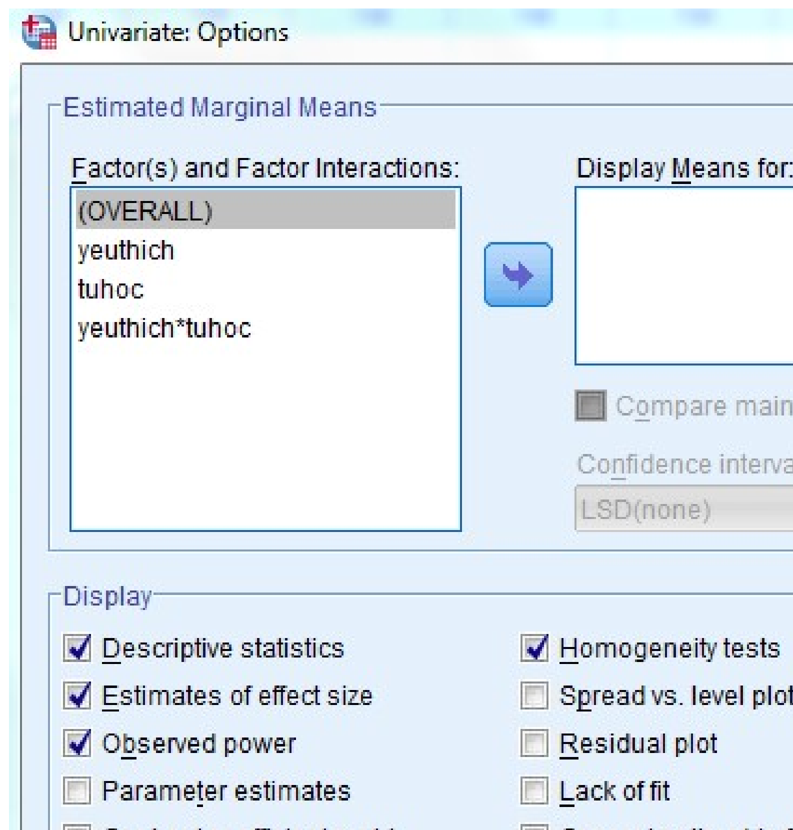
+ ĐTB của SV có mức độ yêu thích ngành đang học khác nhau đều bằng nhau.

+ Không có ảnh hưởng tương tác giữa thời gian tự học và mức độ yêu thích ngành đang học của SV. Một cách cụ thể là, ảnh hưởng của thời gian tự học đến ĐTB là như nhau đối với các nhóm sinh viên có mức độ yêu thích ngành đang học khác nhau, và ảnh hưởng của mức độ yêu thích ngành đang học đến ĐTB là như nhau đối với các nhóm sinh viên có thời gian tự học khác nhau.

B2: Trên menu, chọn Analyze > General Linear Model > Univariate. Đưa biến *diemtb* vào mục **Dependent Variable** và đưa hai biến *yeuthich* va *tuhoc* vào khung **Fixed Factor(s)**.



B3: Chọn nút **Options…**



Nhấn nút **Continue** > **OK.** Ta có các bảng kết quả sau:

Univariate Analysis of Variance

```
[DataSet1] D:\Thanh Phuong\SPSS SV\Phan tich phuong sai
```

Between-Subjects Factors

		Value Label	N
yeuthich	1	khong thich lam	21
	2	thich	21
	3	rat thich	21
tuhoc	1	it	21

Descriptive Statistics

Dependent Variable: diemtb

yeuthich	tuhoc	Mean	Std. Deviation
khong thich lam	it	5.6143	.38483
	trung binh	6.0429	.26367
	nhieu	6.1286	.29277
	Total	5.9286	.37966
thich	it	5.7571	.35989
	trung binh	6.4000	.29439
	nhieu	6.9143	.25448
	Total	6.3571	.56531
rat thich	it	5.7286	.51547
	trung binh	6.7571	.37353
	nhieu	7.3571	.35989
	Total	6.6143	.79705

Levene's Test of Equality of Error Variances[a]

Dependent Variable: diemtb

F	df1	df2	Sig.
1.491	8	54	.182

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + yeuthich + tuhoc + yeuthich * tuhoc

Tests of Between-Subjects Effects

Dependent Variable: diemtb

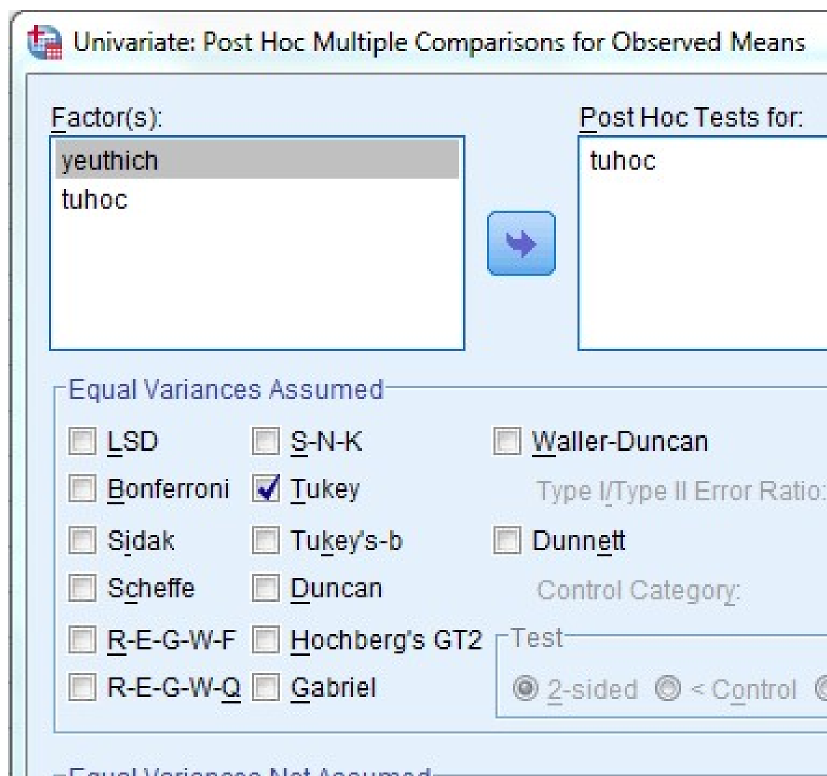
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Nonce Param
Corrected Model	20.306[a]	8	2.538	20.414	.000	.752	16
Intercept	2500.470	1	2500.470	20110.163	.000	.997	2011
yeuthich	5.040	2	2.520	20.267	.000	.429	4
tuhoc	13.020	2	6.510	52.357	.000	.660	10
yeuthich * tuhoc	2.246	4	.561	4.515	.003	.251	1

Hai bảng đầu tiên cung cấp thông tin về các số liệu thống kê mô tả của hai mẫu dữ liệu. Bảng thứ ba là kết quả kiểm định về sự bằng nhau của phương sai các nhóm. Kết quả kiểm định Levene cho thấy giả định phương sai bằng nhau đã không bị vi phạm. Cuối cùng là bảng kết quả kiểm định chính, bảng này cho thấy hai nhân tố chính cũng như sự tương tác

của chúng đều có ảnh hưởng đến kết quả học tập. Tại mức ý nghĩa 5%, các sig. đều bé hơn mức ý nghĩa nên cả ba giả thuyết đặt ra đều bị bác bỏ. Do đó, ta phải tiếp tục phân tích sâu ANOVA. Bằng kiểm định Turkey, chúng ta có bảng kết quả sau cho từng tình huống kiểm định.

a. Xác định các cặp trung bình tổng thể khác nhau theo yếu tố nghiên cứu thứ nhất là thời gian tự học.

Trong hộp thoại Univariate chọn Post Hoc ta được hình sau:



Đưa biến *tuhoc* vào ô **Post Hoc Tests for** và chọn **Tukey.** Nhấn nút **Continue > OK.** Ta có các bảng kết quả sau:

Post Hoc Tests

tuhoc

Multiple Comparisons

Dependent Variable: diemtb
Tukey HSD

(I) tuhoc	(J) tuhoc	Mean Difference (I-J)	Std. Error	Sig.	95% Conf
					Lower Bounc
it	trung binh	-.7000*	.10882	.000	-.962:
	nhieu	-1.1000*	.10882	.000	-1.362:
trung binh	it	.7000*	.10882	.000	.4377
	nhieu	-.4000*	.10882	.002	-.662:
nhieu	it	1.1000*	.10882	.000	.8377

Homogeneous Subsets

diemtb

Tukey HSD[a,b]

tuhoc	N	Subset		
		1	2	3
it	21	5.7000		
trung binh	21		6.4000	
nhieu	21			6.8
Sig.		1.000	1.000	1.

Means for groups in homogeneous subsets are displayed.
Based on observed means

Tất cả cá giá trị Sig. đều bé hơn 0.05 nên ta có thể bác bỏ giả thuyết H0 và kết luận rằng sinh viên mà có thời gian tự học khác nhau sẽ có kết quả học tập khác nhau.

b. Xác định cặp trung bình tổng thể khác nhau theo yếu tố thứ hai là mức độ yêu thích ngành học

Post Hoc Tests

yeuthich

Multiple Comparisons

Dependent Variable: diemtb
Tukey HSD

(I) yeuthich	(J) yeuthich	Mean Difference (I-J)	Std. Error	Sig.	95% Co
					Lower Bou
khong thich lam	thich	-.4286*	.10882	.001	-.69
	rat thich	-.6857*	.10882	.000	-.94
thich	khong thich lam	.4286*	.10882	.001	.16
	rat thich	-.2571	.10882	.056	-.51
rat thich	khong thich lam	.6857*	.10882	.000	.42

Homogeneous Subsets

diemtb

Tukey HSD[a,b]

yeuthich	N	Subset 1	2
khong thich lam	21	5.9286	
thich	21		6.3
rat thich	21		6.6
Sig.		1.000	

Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = .124

Giá trị Sig. cho thấy các sinh viên có mức độ yêu thích ngành học nhiều hoặc rất nhiều thì không có khác biệt nhau về kết quả học tập vì giá trị Sig. cho chênh lệch giữa hai nhóm 2 và nhóm 3 lớn hơn mức ý nghĩa 0.05.

Chương 4

ỨNG DỤNG MÔ HÌNH HỒI QUY TUYẾN TÍNH

4.1. Phân tích hồi quy

Phân tích hồi quy là một trong những công cụ cơ bản của kinh tế lượng. Phân tích hồi quy là mô tả mối quan hệ phụ thuộc của một biến (được gọi là biến phụ thuộc hay biến được giải thích) vào một hay nhiều biến khác (được gọi là biến độc lập hay biến giải thích). Thuật ngữ hồi quy được Francis Galton sử dụng khi ông nghiên cứu các mối quan hệ giữa chiều cao của những đứa trẻ và chiều cao của bố mẹ chúng. Ông thấy rằng mặc dù bố mẹ cao hay thấp thì cũng có những đứa trẻ thấp hay cao, nhưng có một xu thế là chiều cao của những đứa trẻ sẽ hội tụ về một chiều cao trung bình nào đó phụ thuộc một phần vào chiều cao của bố mẹ. Khi số biến độc lập bằng 1 thì ta gọi là hồi quy đơn. Chẳng hạn mô hình hồi quy đơn với một biến phụ thuộc Y và một biến độc lập X trong đó Y là mức chi tiêu và X là thu nhập. Khi số biến độc lập lớn hơn 1 thì ta gọi là hồi quy bội. Chẳng hạn mô hình hồi quy bội với một biến phụ thuộc Y và hai biến độc lập X_1 và X_2 trong đó Y là doanh thu của công ty, X_1 là chi phí cho quảng cáo và X_2 là lương trả cho nhân viên tiếp thị.

Phân tích hồi quy giải quyết những vấn đề sau đây:

- Ước lượng giá trị trung bình của biến phụ thuộc với giá trị đã cho của biến độc lập.
- Dự báo giá trị của Y khi biết được giá trị của biến giải thích X.
- Kiểm định giả thuyết về bản chất của sự phụ thuộc và xác định hiệu quả tác động của biến độc lập lên biến phụ thuộc.

Để hiểu điều này được thực hiện như thế nào, hãy xem xét ví dụ sau:

Ví dụ 1. Giả thiết có một khu phố gồm 50 hộ gia đình. Để nghiên cứu mối quan hệ giữa chi tiêu tiêu dùng hàng tháng của hộ gia đình, ký hiệu là Y (đơn vị: triệu đồng) và thu nhập khả dụng hàng tháng của hộ gia đình hay thu nhập sau khi đã đóng thuế, ký hiệu là X (đơn vị: triệu đồng), chúng ta chia 50 hộ gia đình thành 10 nhóm có thu nhập tương đối như nhau và xem xét chi tiêu tiêu dùng của các hộ gia đình trong từng nhóm thu nhập này.

Bảng 4.1: Thu nhập X và chi tiêu tiêu dùng Y hàng tháng của các hộ gia đình

X	**20**	**25**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**
Y	10	13	12	15	19	20	23	25	30	25
	12	15	13	17	22	22	27	**28**	31	34
	15	19	20	20	24	25	28	30	35	37
	16	21	23	22	27	30	**30**	32	**-**	40
	20	22	25	28	28	33	**32**	35	**-**	**-**

	23	-	27	30	-	-	-	-	-	-
Tổng cộng	**96**	**90**	**120**	**132**	**120**	**130**	**140**	**150**	**96**	**136**

Bảng 4.1 được giải thích như sau: Mỗi cột dọc của Bảng 4.1 cho thấy sự phân phối của chi tiêu tiêu dùng Y ứng với một mức thu nhập X cố định. Chẳng hạn như tương ứng với thu nhập hàng tháng là 20 triệu đồng, có sáu hộ gia đình có mức chi tiêu tiêu dùng hàng tháng trong khoảng 10 đến 23 triệu đồng. Lưu ý rằng các dữ liệu trong Bảng 4.1 tiêu biểu cho tổng thể, do đó chúng ta có thể tính các **xác suất có điều kiện** của Y theo X, kí hiệu là p(Y|X). Điều này có nghĩa là chúng ta thấy được **phân phối có điều kiện** của Y phụ thuộc vào các giá trị nhất định của X. Xác suất có điều kiện của các dữ liệu trong Bảng 4.1 được trình bày trong bảng sau:

Bảng 4.2: Xác suất có điều kiện p(Y|X) của dữ liệu trong Bảng 4.1

X	**20**	**25**	**30**	**35**	**40**	**45**	**50**	**55**	**60**	**65**
P(Y|X)	1/6	1/5	1/6	1/6	1/5	1/5	1/5	1/5	1/3	1/4
	1/6	1/5	1/6	1/6	1/5	1/5	1/5	1/5	1/3	1/4
	1/6	1/5	1/6	1/6	1/5	1/5	1/5	1/5	1/3	1/4
	1/6	1/5	1/6	1/6	1/5	1/5	1/5	1/5		1/4
	1/6	1/5	1/6	1/6	1/5	1/5	1/5	1/5		
	1/6		1/6	1/6						
Trung bình có điều kiện của Y	**16**	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**32**	**34**

Ở bảng trên, đối với mỗi phân phối xác suất có điều kiện của **Y** chúng ta có thể tính được giá trị trung bình của nó, được gọi là *trung bình có điều kiện* hay *kỳ vọng có điều kiện*, được thể hiện bằng $\mathbf{E(Y \mid X = X_i)}$ và được diễn giải là "giá trị kỳ vọng của Y khi X nhận một giá trị cụ thể $\mathbf{X_i}$", để đơn giản hóa về mặt ký hiệu chúng ta viết lại thành $\mathbf{E(Y \mid X_i)}$. Chẳng hạn như $E(Y \mid X = 20) = 10\ (1/6) + 12(1/6) + 15(1/6) + 16(1/6) + 20(1/6) + 23(1/6)$ = **15.8**. Các trung bình có điều kiện khác được tính tương tự và các kết quả này được đặt ở hàng cuối cùng của Bảng 4.2

Nhận xét rằng mặc dù có sự biến đổi trong chi tiêu tiêu dùng của từng hộ gia đình, nhưng chi tiêu tiêu dùng về mặt **trung bình** sẽ tăng khi thu nhập tăng. Về mặt hình học, đồ

thị phân tán cũng cho chúng ta thấy được điều này. Đồ thị phân tán cho thấy rằng các trung bình có điều kiện này nằm trên một đường thẳng. Đường thẳng này được gọi là *đường hồi quy tổng thể.*

4.2. Hàm hồi quy tổng thể

Vì E(Y|X) là một hàm của biến giải thích X nên ta viết E(Y|X) = f (X). Phương trình này được gọi là hàm hồi quy tổng thể (PRF: Population Regression Function) hay hồi quy tổng thể (PR: Population Regression). Hàm f (X) có dạng như thế nào? Trên thực tế, chúng ta không thể có toàn bộ dữ liệu của tổng thể, do đó, dạng hàm của PRF là một vấn đề thực nghiệm. Giả sử rằng chi tiêu tiêu dùng là có quan hệ tuyến tính với thu nhập. Khi đó, **E(Y|X)** là một hàm tuyến tính của **X** và được viết như sau:

$$E(Y \mid X) = \beta_1 + \beta_2 X \tag{4.1}$$

trong đó β_1, β_2 là các tham số không biết nhưng không thể thay đổi được và được gọi là các hệ số hồi quy; β_1 còn được gọi là **hệ số tung độ gốc** hay **hệ số chặn** và β_2 còn được gọi là **hệ số độ dốc** hay **hệ số góc**, Y là biến phụ thuộc hay biến được giải thích (explained variable), X là biến độc lập hay biến giải thích (explanatory variable).

Phương trình (2.1) được gọi là **hàm hồi quy** tổng thể tuyến tính hay **mô hình hồi quy** tổng thể tuyến tính hay **phương trình hồi quy** tổng thể tuyến tính. Trong các phần tiếp theo, các thuật ngữ hàm hồi quy, phương trình hồi quy, và mô hình hồi quy sẽ được dùng với nghĩa như nhau. Thuật ngữ "tuyến tính" ở đây được hiểu như sau: $E(Y \mid X_i)$ là một hàm tuyến tính theo các tham số β_i; nó có thể tuyến tính hoặc có thể không tuyến tính theo biến **X**. Theo cách giải thích này, $E(Y \mid X) = \beta_1 + \beta_2 X$ là một mô hình tuyến tính nhưng $E(Y \mid X) = \beta_1 + \sqrt{\beta_2} X$ thì không phải vì tham số β_2 có lũy thừa bằng 1/2. Trong phần này, chúng ta sẽ không bàn tới những mô hình hồi quy không tuyến tính theo các tham số.

Từ Bảng 4.2, nhận thấy rằng khi thu nhập hàng tháng của các hộ gia đình tăng, chi tiêu tiêu dùng của các hộ gia đình về mặt trung bình cũng tăng theo. Tuy nhiên, chi tiêu tiêu dùng của từng hộ gia đình không nhất thiết phải tăng khi mức thu nhập tăng. Ví dụ, trong Bảng 4.1 chúng ta quan sát thấy tương ứng với mức thu nhập 55 triệu đồng một tháng có một gia đình với mức chi tiêu tiêu dùng là 28 triệu đồng, thấp hơn mức chi tiêu tiêu dùng của hai hộ gia đình mà mức thu nhập hàng tháng của họ là 50 triệu đồng. Nhưng lưu ý rằng mức chi tiêu tiêu dùng **trung bình** của các hộ gia đình với thu nhập hàng tháng là 55 triệu đồng là lớn hơn mức chi tiêu tiêu dùng trung bình của những hộ gia đình có mức thu nhập hàng tháng là 50 triệu đồng (28 triệu đồng so với 30 triệu đồng). Như vậy, chúng ta có thể nói gì về mối tương quan giữa mức chi tiêu tiêu dùng của từng hộ gia đình và một mức thu nhập nhất định. Chúng ta thấy rằng với mức thu nhập là X_i, mức chi tiêu tiêu dùng của từng hộ gia đình xoay xung quanh chi tiêu trung bình của tất cả các hộ gia

đình ở tại **X_i**, có nghĩa là xung quanh kỳ vọng có điều kiện của nó. Do đó, chúng ta có thể diễn đạt **độ lệch** của một giá trị **Y_i** xung quanh giá trị kỳ vọng của nó như sau:

$$u_i = Y_i - E(Y \mid X_i)$$

hay

$$Y_i = E(Y \mid X_i) + u_i \quad (4.2)$$

trong đó độ lệch **u_i** là một biến số ngẫu nhiên không thể quan sát và nhận giá trị âm hoặc dương. Đại lượng u_i còn được gọi là **số hạng nhiễu ngẫu nhiên** hay **số hạng sai số ngẫu nhiên.**

Từ phương trình (4.2), chúng ta có thể nói rằng với một mức thu nhập cụ thể, chi tiêu của một hộ gia đình được thể hiện như là tổng của hai thành tố. Thành tố thứ nhất là chi tiêu tiêu dùng trung bình của tất cả các hộ gia đình có cùng mức thu nhập $E(Y \mid X_i)$. Thành tố này được gọi là **thành tố tất định hay hệ thống**. Thành tố thứ hai là **thành tố ngẫu nhiên hay không hệ thống**. Chúng ta có thể giả định rằng thành tố ngẫu nhiên là một số hạng thay thế cho tất cả các biến số ta bỏ ra ngoài hay bỏ sót mà có thể ảnh hưởng đến **Y** nhưng không đưa vào trong mô hình hồi quy. Nếu $E(Y \mid X_i)$ được giả định là tuyến tính theo **X_i** thì phương trình (4.2) được biểu diễn như sau:

$$Y_i = E(Y \mid X_i) + u_i = \beta_1 + \beta_2 X_i + u_i \quad (4.3)$$

Phương trình (4.3) giả định rằng chi tiêu tiêu dùng của một hộ gia đình có quan hệ tuyến tính đối với thu nhập cộng với số hạng nhiễu. Như vậy, chi tiêu tiêu dùng của một hộ gia đình, với thu nhập hàng tháng X = 55 triệu đồng, có thể được biểu diễn như sau:

$$Y_1 = 25 = \beta_1 + \beta_2 \times 55 + u_1$$
$$Y_2 = 28 = \beta_1 + \beta_2 \times 55 + u_2$$
$$Y_3 = 30 = \beta_1 + \beta_2 \times 55 + u_3$$
$$Y_4 = 32 = \beta_1 + \beta_2 \times 55 + u_4$$
$$Y_5 = 35 = \beta_1 + \beta_2 \times 55 + u_5$$

Lấy kỳ vọng hai vế của phương trình (4.2), ta được

$$E(Y_i \mid X_i) = E\left[E(Y \mid X_i)\right] + E(u_i \mid X_i)$$
$$\Leftrightarrow E(Y_i \mid X_i) = E(Y \mid X_i) + E(u_i \mid X_i) \quad (4.4)$$

Vì $E(Y_i \mid X_i) = E(Y \mid X_i)$ nên $E(u_i \mid X_i) = 0.$ Như vậy, giả định cho rằng đường hồi quy đi ngang qua các giá trị trung bình có điều kiện của **Y** có nghĩa là các giá trị trung bình có điều kiện của **u_i** (phụ thuộc vào các giá trị của **X**) là bằng 0.

4.3. Hàm hồi quy mẫu

Giả sử rằng chúng ta chưa biết được thông tin của tổng thể được cho ở Bảng 4.1 và chúng ta chỉ có thông tin của hai mẫu ngẫu nhiên như ở Bảng 4.3 và Bảng 4.4. Trong Bảng 4.3 hay Bảng 4.4, ứng với mỗi giá trị của X, chỉ có một giá trị duy nhất Y.

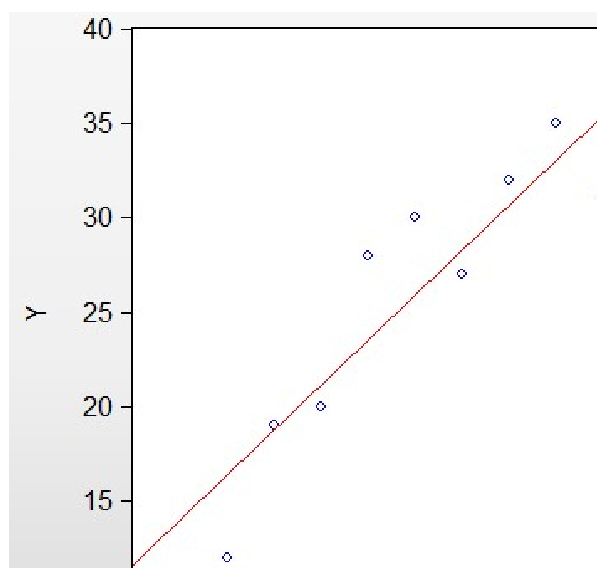
Bảng 4.3: Mẫu ngẫu nhiên thứ nhất từ tổng thể

Thu nhập X	20	25	30	35	40	45	50	55	60	65
Chi tiêu Y	12	19	20	28	30	27	32	35	31	37

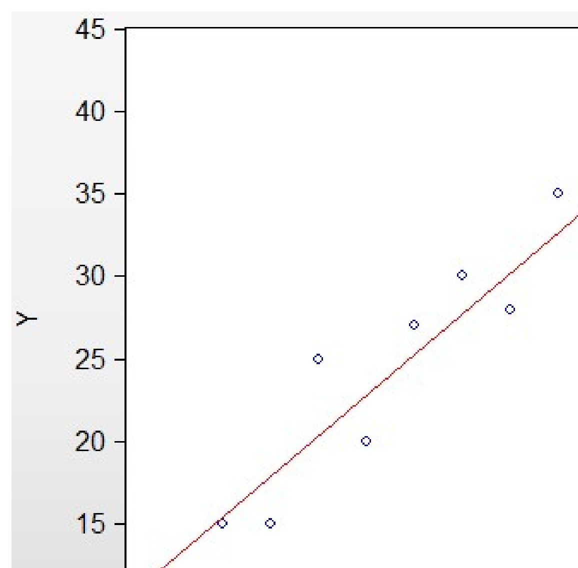
Bảng 4.4: Mẫu ngẫu nhiên thứ hai từ tổng thể

Thu nhập X	20	25	30	35	40	45	50	55	60	65
Chi tiêu Y	15	15	25	20	27	30	28	35	30	40

Dữ liệu ở Bảng 4.3 và Bảng 4.4 được thể hiện ở hình sau:



Hình 4.2: Dữ liệu từ Bảng 4.3



Hình 4.3: Dữ liệu từ Bảng 4.4

Ứng với mỗi mẫu, sẽ có một đường hồi quy tương ứng. Đường hồi quy này được gọi là **đường hồi quy mẫu**. Nhận xét rằng với bao nhiêu mẫu ngẫu nhiên thì sẽ có bấy nhiêu đường hồi quy mẫu. Những đường hồi quy mẫu này nói chung không giống nhau và phản ánh được tính chất của đường hồi quy tổng thể. Tương ứng với đường hồi quy mẫu, chúng ta thiết lập một hàm số, được gọi là hàm hồi quy mẫu (Sample Regression Function-SRF). Hàm hồi quy mẫu ứng với hàm hồi quy tổng thể

$$E(Y \mid X_i) = \beta_1 + \beta_2 X_i$$

có dạng như sau:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \tag{4.5}$$

trong đó, $\hat{Y}_i$ là hàm ước lượng của $E(Y|X_i)$, $\hat{\beta}_1$ và $\hat{\beta}_2$ được gọi là các hệ số hồi quy mẫu và lần lượt là các ước lượng của các hệ số hồi quy tổng thể β_1 và β_2 tương ứng.

Hàm hồi quy mẫu cũng có thể được biểu diễn dưới dạng ngẫu nhiên cho từng quan sát:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \tag{4.6}$$

hay một cách tổng quát:

$$Y = \hat{\beta}_1 + \hat{\beta}_2 X + \hat{u} \tag{4.7}$$

trong đó $\hat{u}_i$ là số hạng phần dư mẫu và $\hat{u}_i$ có thể được xem là một ước lượng của u_i. Việc đưa thành phần $\hat{u}_i$ vào trong hàm hồi quy mẫu và việc đưa thành phần u_i vào trong hàm hồi quy tổng thể là có cùng lý do.

Câu hỏi đặt ra là làm thế nào để ước lượng chính xác tối đa hàm hồi quy tổng thể (PRF) trên cơ sở hàm hồi quy mẫu (SRF). Để thực hiện điều này, **phương pháp bình phương tối thiểu thông thường (OLS)** là một trong những phương pháp được sử dụng phổ biến nhất.

4.4. Phương pháp bình phương tối thiểu thông thường (Ordinary Least Squared-OLS)

4.4.1. Nội dung phương pháp OLS

Xét hàm hồi quy tổng thể dưới dạng tuyến tính:

$$Y = \beta_1 + \beta_2 X + U$$

Ta cần ước lượng các hệ số β_1, β_2. Giả sử có mẫu ngẫu nhiên kích thước n là {(X_i, Y_i), i = 1, …, n} được rút ra từ tổng thể. Từ mẫu này, ước lượng các giá trị của β_1, β_2, ta được hàm hồi quy mẫu sau:

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$$

hay có thể viết cho từng quan sát

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

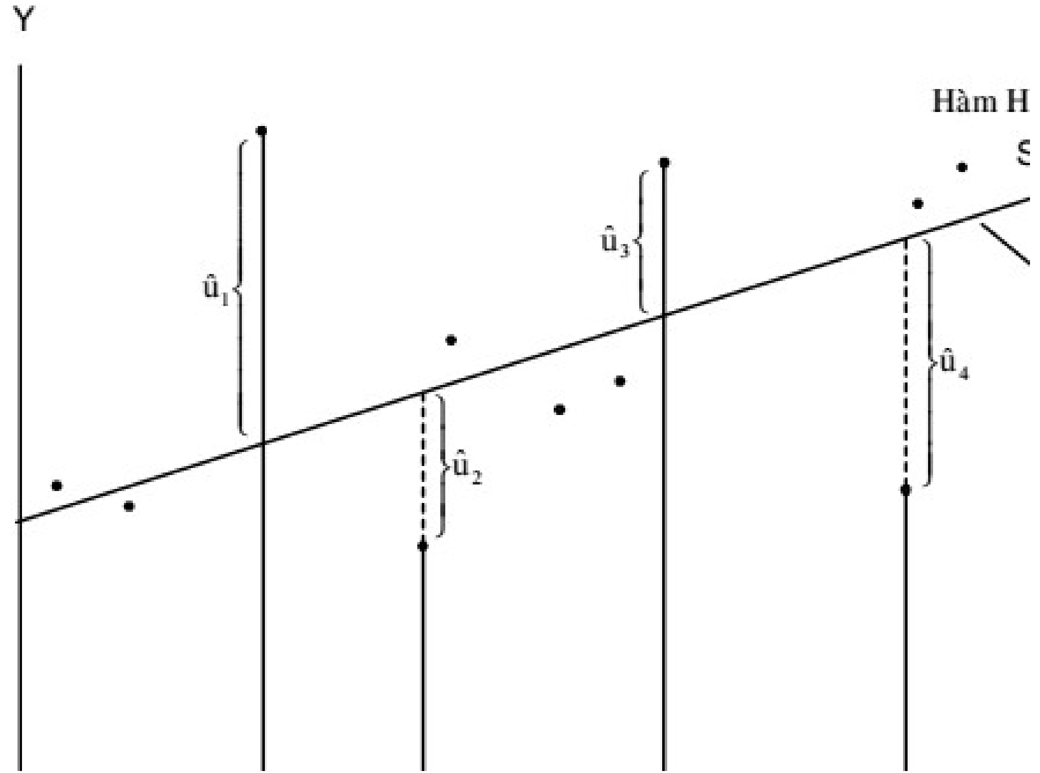
Các dạng trên là dạng xác định của hàm hồi quy mẫu, ta có thể viết hàm hồi quy mẫu dưới dạng ngẫu nhiên như sau:

$$Y = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{U}$$

hay có thể viết cho từng quan sát

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \widehat{U}_i$$

Đặt $\widehat{U}_i = Y_i - \hat{Y}_i$ và đại lượng này được gọi là phần dư (Residuals). Hàm hồi quy mẫu và phần dư có thể được minh họa như hình dưới đây:



Hình 4.4: Hàm hồi quy mẫu và phần dư

Ta muốn xác định các giá trị $\hat{\beta}_1, \hat{\beta}_2$ sao cho tổng bình phương các sai lệch giữa các giá trị thực tế Y_i và giá trị ước lượng tương ứng từ hàm hồi quy mẫu là nhỏ nhất, tức là tìm các giá

trị $\hat{\beta}_1, \hat{\beta}_2$ sao cho $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_i)^2 = f(\widehat{\beta}_1, \widehat{\beta}_2) \to \min$

Đây là bài toán cực trị hai biến không có điều kiện ràng buộc, do đó $\widehat{\beta_1}$, $\widehat{\beta_2}$ sẽ là nghiệm của hệ phương trình sau:

$$\begin{cases} \dfrac{\partial f}{\partial \hat{\beta}_1} = -2\sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0 \\ \dfrac{\partial f}{\partial \hat{\beta}_2} = -2\sum X_i (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0 \end{cases} \Rightarrow \begin{cases} \hat{\beta}_1 n + \hat{\beta}_2 \sum X_i = \sum Y_i \\ \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 = \sum X_i Y_i \end{cases}$$

Đặt $\overline{X} = (\sum X_i)/n$; $\overline{Y} = (\sum Y_i)/n$; $\overline{X^2} = (\sum X_i^2)/n$, $x_i = X_i - \overline{X}$; $y_i = Y_i - \overline{Y}$.

Khi đó,

$$\begin{cases} \hat{\beta}_2 = \dfrac{\sum\limits_{i=1}^{n} x_i y_i}{\sum\limits_{i=1}^{n} x_i^2} = \dfrac{\sum\limits_{i=1}^{n} X_i Y_i - n\overline{X}\overline{Y}}{\sum X_i^2 - n(\overline{X})^2} \\ \hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X} \end{cases}$$

Ta gọi $\hat{\beta}_1, \hat{\beta}_2$ là các ước lượng bình phương nhỏ nhất (OLS) của β_1, β_2.

Ví dụ 2. Bảng sau đây cho số liệu về mức chi tiêu tiêu dùng hàng tháng của hộ gia đình, ký hiệu là Y (đơn vị: triệu đồng) và thu nhập khả dụng hàng tháng của hộ gia đình hay thu nhập sau khi đã đóng thuế, ký hiệu là X (đơn vị: triệu đồng) của một mẫu gồm 10 hộ gia đình. Giả sử Y và X có mối quan hệ tuyến tính. Hãy ước lượng hàm hồi quy của Y theo X.

Bảng 4.5: Mẫu ngẫu nhiên từ tổng thể

Thu nhập X	20	25	30	35	40	45	50	55	60	65
Chi tiêu Y	12	19	20	28	30	27	32	35	31	37

Từ các số liệu quan sát của X và Y ở Bảng 2.5, ta tính được

$$\sum X_i = 425; \sum Y_i = 265; \sum X_i^2 = 20125; \sum X_i Y_i = 12280; \overline{X} = 42.5; \overline{Y} = 26.5$$

Do đó,

$$\begin{cases} \hat{\beta}_2 = \dfrac{\sum\limits_{i=1}^{n} x_i y_i}{\sum\limits_{i=1}^{n} x_i^2} = \dfrac{\sum\limits_{i=1}^{n} X_i Y_i - n\overline{X}\overline{Y}}{\sum X_i^2 - n(\overline{X})^2} = \dfrac{12280 - 10 \times 42.5 \times 26.5}{20125 - 10 \times (42.5)^2} = \dfrac{1017.5}{2062.5} = 0.493 \\ \hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X} = 26.5 - 0.493 \times 42.5 = 5.5475 \end{cases}$$

Vậy hàm hồi quy mẫu của chi tiêu tiêu dùng theo mức thu nhập là:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i = 5.5475 + 0.493 X_i$$

Ý nghĩa các hệ số ước lượng:

- $\hat{\beta}_1 = 5.5475$: Chi tiêu tiêu dùng trung bình hàng tháng của một hộ gia đình khi thu nhập hàng tháng bằng 0.
- $\hat{\beta}_2 = 0.493$: Xét mức thu nhập nằm trong khoảng (20, 65) triệu đồng một tháng, nếu khi thu nhập tăng thêm 1 triệu đồng/tháng thì chi tiêu tiêu dùng trung bình tăng thêm khoảng 0.493 triệu đồng/tháng.

Thực hiện trên Eviews ta được kết quả sau:

Dependent Variable: Y

Method: Least Squares
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X	0.493333	0.073251	6.734867	0.0001
C	5.533333	3.286089	1.683866	0.1307

R-squared	0.850071	Mean dependent var	26.50000
Adjusted R-squared	0.831329	S.D. dependent var	8.100069
S.E. of regression	3.326660	Akaike info criterion	5.418671
Sum squared resid	88.53333	Schwarz criterion	5.479188
Log likelihood	-25.09336	Hannan-Quinn criter.	5.352284
F-statistic	45.35843	Durbin-Watson stat	3.303012
Prob(F-statistic)	0.000147		

Chú thích cho bảng kết quả như sau:

- *Dependent Variable*: Biến phụ thuộc Y.
- *Method: Least Squares*: Sử dụng phương pháp bình phương tối thiểu OLS.
- *Date, Time*: Ngày, giờ thực hiện.
- *Sample*: Phạm vi của mẫu quan sát.
- *Included observations*: Tổng số quan sát (cỡ mẫu).
- *Variable*: Danh sách các biến độc lập trong mô hình hồi quy, trong đó C là hệ số β_1.
- *Coefficient:* Các ước lượng hệ số của mô hình.
- *Std. Error*: Sai số chuẩn của $\widehat{\beta}_1$ và $\widehat{\beta}_2$.
- *t-Statistic*: Giá trị (quan sát) của thống kê T.
- *Prob.*: Giá trị p-value của thống kê T.
- *R-squared*: Hệ số xác định (hệ số tương quan toàn phần) R^2.
- *Adjusted R-squared*: Hệ số xác định đã được hiệu chỉnh $\overline{R}^2$
- *S.E. of regression*: Sai số tiêu chuẩn của hàm hồi quy.
- *Sum squared resid*: Tổng bình phương sai số *RSS*.
- *Log likelihood*: Logarit cơ số e của hàm hợp lý.
- *F-statistic*: Giá trị thống kê của thống kê F.

- *Prob(F-statistic)*: $P(F > F - statistic)$.
- *Mean dependent var*: Trung bình của biến phụ thuộc.
- *S.D. dependent var*: Độ lệch chuẩn của biến phụ thuộc.
- *Akaike info criterion*: Tiêu chuẩn Akaike.
- *Schwarz criterion*: Tiêu chuẩn Schwarz.
- *Hannan-Quinn criter.*: Tiêu chuẩn Hannan-Quinn.
- *Durbin-Watson stat*: Thống kê Durbin-Watson.

4.4.2. Mô hình hồi quy tuyến tính cổ điển: Các giả thiết cơ bản của phương pháp OLS

Mô hình hồi quy tuyến tính cổ điển hay mô hình chuẩn, mô hình Gauss (CLRM) được coi là nền tảng của hầu hết lý thuyết kinh tế lượng, gồm 10 giả thiết sau:

Giả thiết 1: Mô hình hồi quy là tuyến tính theo các tham số, tức là

$$Y_i = \beta_1 + \beta_2 X_i + U_i \tag{4.8}$$

Giả thiết 2: Các giá trị X được cố định trong việc lấy mẫu lập lại. Điều này có nghĩa là X được giả thiết là không ngẫu nhiên.

Chẳng hạn như với dữ liệu ở Bảng 2.1, để có các mẫu ngẫu nhiên, ta giữ giá trị thu nhập X cố định và giả sử bằng 55 triệu đồng, ta rút ra một cách ngẫu nhiên một hộ gia đình nào đó và có được thông tin về chi tiêu hàng tháng **Y** của hộ gia đình đó, giả sử là 28 triệu đồng. Vẫn giữ X ở mức 55 triệu đồng, ta lại rút một cách ngẫu nhiên một hộ gia đình khác và thấy giá trị quan sát Y của nó là 32 triệu đồng. Trong mỗi lần rút ra một hộ gia đình để xem xét, giá trị X được cố định ở mức 55 triệu đồng. Ta có thể lặp lại quá trình này cho tất cả các giá trị X đã ghi trong Bảng 2.1. Tất cả những điều này có nghĩa là sự phân tích hồi quy được trình bày là *phân tích hồi quy có điều kiện*, nghĩa là các giá trị của (các) biến hồi quy độc lập X được cho trước.

Giả thiết 3: Cho trước giá trị của X, giá trị trung bình hay kỳ vọng của các số hạng nhiễu U_i bằng 0, tức là

$$E(U_i \mid X_i) = 0.$$

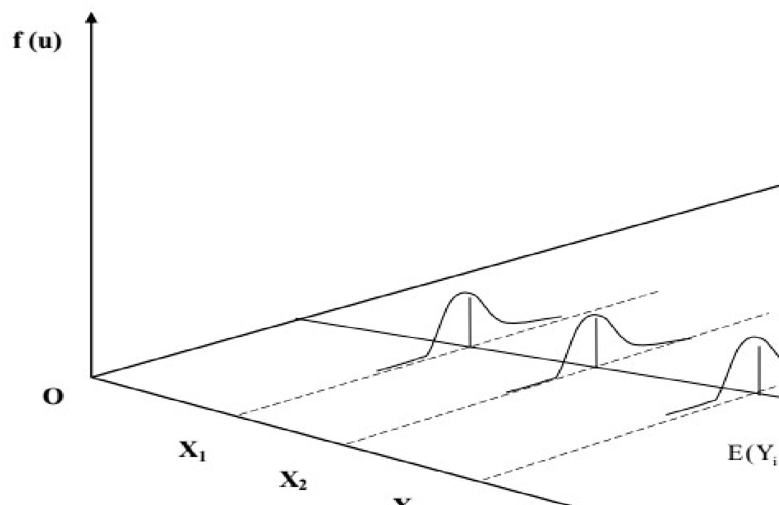
Nhận xét rằng giả thiết này ngụ ý rằng $E(Y \mid X_i) = \beta_1 + \beta_2 X_i$. Do đó, hai giả thiết này là tương đương với nhau.

Giả thiết 4: Cho các giá trị của X, phương sai của U_i sẽ như nhau đối với tất cả mọi quan sát, tức là:

$$\mathrm{var}(U_i \mid X_i) = E\left[U_i - E(U_i)\middle|X_i\right]^2 = E\left(U_i^2 \mid X_i\right) = \sigma^2, \forall i = \overline{1,n}$$

trong đó var là ký hiệu của phương sai.

Về mặt đồ thị, giả thiết này được mô tả như hình dưới đây:



Hình 4.5: Phương sai có điều kiện không đổi

Giả thiết 5: Cho trước hai giá trị bất kỳ X_i và X_j với $i \neq j$, tương quan giữa U_i và U_j bất kỳ với $i \neq j$ bằng 0, tức là

$$\text{cov}\left(U_i, U_j \mid X_i, X_j\right) = E\left[\left(U_i - E(U_i) \mid X_i\right)\left(U_j - E(U_j) \mid X_j\right)\right] = E\left[\left(U_i \mid X_i\right)\left(U_j \mid X_j\right)\right] = 0$$

trong đó i và j là hai quan sát khác nhau và cov là ký hiệu của đồng phương sai.

Giả thiết 6: Đồng phương sai giữa U_i và X_i bằng 0. Ký hiệu là cov (U_i, X_i) = 0.

Giả thiết 7: Số lượng các quan sát n phải lớn hơn số lượng các biến giải thích.

Giả thiết 8: Các giá trị X trong một mẫu cho trước không thể tất cả đều bằng nhau.

Giả thiết 9: Mô hình hồi quy được xác định một cách đúng đắn. Điều này có nghĩa là phải xác định đúng các biến xuất hiện trong mô hình, dạng hàm hồi quy, quy luật phân phối xác suất của các biến ngẫu nhiên.

Giả thiết 10: Không có các mối tương quan tuyến tính hoàn hảo trong các biến giải thích.

4.4.3. Các tính chất của hàm ước lượng OLS

Các tính chất của hàm ước lượng OLS được thể hiện ở Định lý Gaus-Markov, cụ thể như sau:

Định lý Gauss-Markov: Cho trước các giả thiết của mô hình hồi quy tuyến tính cổ điển, các
hàm ước lượng bình phương tối thiểu, trong nhóm các hàm ước lượng tuyến tính không chệch, có phương sai nhỏ nhất, nghĩa là chúng là các hàm ước lượng không chệch tuyến tính
tốt nhất (BLUE: Best Linear Unbiased Estimation).

Tính tuyến tính của các hàm ước lượng $\hat{\beta}_1, \hat{\beta}_2$ được hiểu là $\hat{\beta}_1, \hat{\beta}_2$ là các hàm tuyến tính theo các biến ngẫu nhiên $Y_1, Y_2, \ldots, Y_n$.

Các hàm ước lượng $\hat{\beta}_1, \hat{\beta}_2$ không chệch, có nghĩa là $E\left(\widehat{\beta_1}\right) = \beta_1, E\left(\widehat{\beta_2}\right) = \beta_2$.

Các hàm ước lượng $\hat{\beta}_1, \hat{\beta}_2$ có phương sai nhỏ nhất trong nhóm tất cả các hàm ước lượng không chệch tuyến tính; hàm ước lượng không chệch với phương sai tối thiểu được gọi là **hàm ước lượng hiệu quả.**

4.4.4. Độ chính xác của các ước lượng OLS

Tính không chệch của ước lượng cho biết trung bình của sai lệch của các giá trị $\hat{\beta}_j$ thu được từ các mẫu khác nhau so với β_j bằng 0. Tuy nhiên, ta lại không biết sai lệch này có thể lớn thế nào. Nếu các sai lệnh này là nhỏ thì khi lấy một mẫu bất kỳ, ta có thể hy vọng rằng giá trị $\hat{\beta}_j$ thu được là không quá khác biệt so với giá trị β_j chưa biết. Khi đó, ta nói độ chính xác của $\hat{\beta}_j$ là cao. Như vậy, độ chính xác của các ước lượng được đo bởi độ phân tán của $\hat{\beta}_j$ xung quanh β_j tương ứng, nghĩa là $E\left[\left(\widehat{\beta}_j - \beta_j\right)^2\right]$. Khi $\hat{\beta}_j$ là ước lượng không chệch của β_j, nghĩa là $E\left(\widehat{\beta}_j\right) = \beta_j$, thì độ chính xác này chính là phương sai của các ước lượng. Cụ thể như sau:

$$E\left[\left(\widehat{\beta}_j - \beta_j\right)^2\right] = E\left\{\left[\widehat{\beta}_j - E\left(\beta_j\right)\right]^2\right\} = \text{var}\left(\widehat{\beta}_j\right)$$

Khi phương sai càng bé thì độ chính xác của ước lượng càng cao. Phương sai của các ước lượng được thể hiện qua định lý dưới đây.

Định lý: *Khi các giả thiết 1, giả thiết 2, giả thiết 3 và giả thiết 4 được thỏa mãn thì phương sai của các hệ số ước lượng được xác định bởi*

$$var\left(\widehat{\beta_1}\right) = \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\sigma^2$$

$$var\left(\widehat{\beta_2}\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

Trong thực tế ta thường không biết σ^2, do đó trong các công thức trên, ta thay σ^2 bởi ước lượng điểm không chệch, tốt nhất của nó, kí hiệu bởi $\widehat{\sigma^2}$ và được xác định như sau:

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n} \widehat{U}_i^2}{n-2}$$

Đại lượng $\widehat{\sigma}$ còn được gọi là sai số chuẩn của hàm hồi quy (Standard Error of Regression).

Như vậy, các sai số chuẩn (Standard error) của các hệ số ước lượng $\widehat{\beta}_1$, $\widehat{\beta}_2$ được xác định bởi các công thức sau:

$$se\left(\widehat{\beta}_1\right) = \sqrt{var\left(\widehat{\beta}_1\right)} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\widehat{\sigma^2}}$$

$$se\left(\widehat{\beta}_2\right) = \sqrt{var\left(\widehat{\beta}_2\right)} = \sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}}$$

4.5. Độ phù hợp của hàm hồi quy mẫu. Hệ số xác định R^2

Để xác định sự thay đổi của biến độc lập giải thích được bao nhiêu phần trăm sự thay đổi của biến phụ thuộc, chúng ta sử dụng một đại lượng được gọi là hệ số xác định (Coefficient of Determination), ký hiệu là r^2 trong trường hợp hồi quy hai biến và ký hiệu là R^2 trong trường hợp hồi quy đa biến. Phương pháp tính r^2 sẽ được trình bày ở phần dưới đây.

Ta có $Y_i = \widehat{Y}_i + \widehat{U}_i$ hay $y_i = \hat{y}_i + \hat{u}_i$. Bình phương hai vế của phương trình này, lấy tổng đối với mẫu và sử dụng các kết quả $\sum \hat{y}_i \hat{u}_i = 0$ và $\hat{y}_i = \hat{\beta}_2 x_i$, ta được:

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2\sum \hat{y}_i \hat{u}_i = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 = \hat{\beta}_2^2 \sum x_i^2 + \sum \hat{u}_i^2$$

Trong đó, tổng $\sum y_i^2 = \sum\left(Y_i - \overline{Y}\right)^2$ là độ lệch tổng cộng các giá trị thực của Y so với trung bình mẫu của chúng và được gọi là tổng bình phương toàn phần (TSS: Total Sum of Squares); tổng $\sum \hat{y}_i^2 = \sum\left(\widehat{Y}_i - \overline{\widehat{Y}}\right)^2 = \sum\left(\widehat{Y}_i - \overline{Y}\right)^2 = \hat{\beta}_2^2 \sum x_i^2$ là chênh lệch của giá trị ước lượng của Y với trung bình của chúng và được gọi là tổng bình phương hồi quy (ESS: Explained Sum of Squares). Lưu ý rằng $\overline{\widehat{Y}} = \overline{Y}$. Tổng $\sum \hat{u}_i^2$ được gọi là tổng bình phương phần dư hay tổng bình phương sai số (RSS: Residual Sum of Squares). Do đó, ta có:

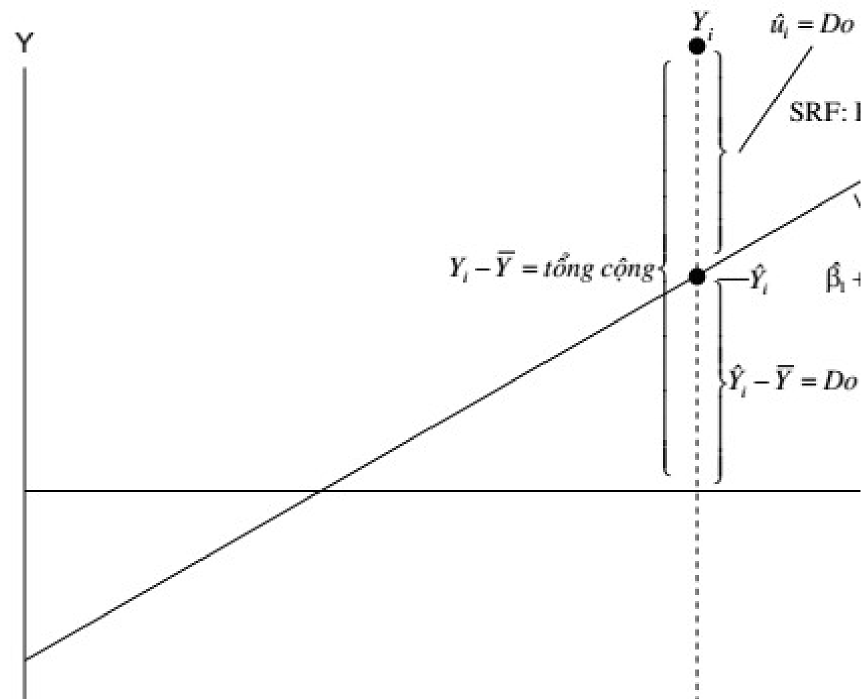
TSS = ESS + RSS

Chia hai vế của phương trình trên ta được

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS} = \frac{\sum\left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2} + \frac{\sum \hat{u}_i^2}{\sum\left(Y_i - \overline{Y}\right)^2}$$

Đại lượng r^2 được xác định bởi tỷ số $\frac{ESS}{TSS}$, biểu thị phần trăm của độ lệch tổng cộng trong Y được giải thích bởi mô hình hồi quy. Ta viết

$$r^2 = \frac{ESS}{TSS} = \frac{\sum\left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\sum \hat{u}_i^2}{\sum\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{RSS}{TSS}$$



Hình 4.6: Minh họa tổng bình phương phần dư

Chương 5

MÔ HÌNH HỒI QUY BỘI

Trong thực tế, các mối quan hệ kinh tế thường phức tạp, một biến số kinh tế có thể chịu sự tác động của nhiều biến số kinh tế khác nhau. Chẳng hạn, khi nghiên cứu nhu cầu về một loại hàng hóa nào đó thì nhu cầu này phụ thuộc đồng thời vào nhiều yếu tố như thu nhập của người tiêu dùng, giá bán của hàng hóa đó, thị hiếu người tiêu dùng… Do đó, cần thiết phải nghiên cứu mô hình hồi quy nhiều hơn hai biến, còn gọi là mô hình hồi quy bội (Multiple regression).

5.1. Mô hình hồi quy bội tuyến tính

Mô hình hồi quy tuyến tính k biến có thể viết dưới dạng sau:

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + U \qquad (5.1)$$

trong đó Y là biến phụ thuộc, $X_j, j = \overline{2,k}$ là các biến độc lập, $\beta_j, j = \overline{1,k}$ được gọi là các hệ số hồi quy bội, U là sai số ngẫu nhiên, đại diện cho các yếu tố khác ngoài X_j có tác động đến Y nhưng không được đưa vào mô hình với lý do chúng ta không có quan sát về nó, hoặc không muốn đưa nó vào mô hình, hoặc không thể đưa nó vào mô hình.

Các giả thiết cho mô hình (5.1):

Giả thiết 1: *Mô hình được ước lượng dựa trên mẫu ngẫu nhiên*

$$\left\{(X_{ji}, Y_i) : i = \overline{1,n}, j = \overline{2,k}\right\}$$

Giả thiết 2: *Kỳ vọng của sai số ngẫu nhiên tại mỗi giá trị ($X_{2i}, X_{3i}, \ldots, X_{ki}$) bằng 0*

$$E(U \mid X_{2i}, \ldots, X_{ki}) = 0, \forall i = \overline{1,n}$$

Giả thiết 3: *Phương sai của sai số ngẫu nhiên tại các giá trị ($X_{2i}, X_{3i}, \ldots, X_{ki}$) là hằng số*

$$Var(U \mid X_{2i}, \ldots, X_{ki}) = \sigma^2, \forall i = \overline{1,n}$$

Giả thiết 4: *Giữa các biến độc lập $X_j, j = \overline{2,k}$ không có mối quan hệ đa cộng tuyến hoàn hảo, có nghĩa là không tồn tại các hằng số $\lambda_j, j = \overline{2,k}$ không đồng thời bằng 0 sao cho* $\sum_{j=2}^{k} \lambda_j X_j = 0$.

Ví dụ 1. Để xem tác động của các hình thức đầu tư lên GDP, người ta sử dụng hàm hồi quy bội tuyến tính

$$GDP = \beta_1 + \beta_2 GI + \beta_3 PI + \beta_4 FDI + \beta_5 I + U$$

với GI, DI, FDI, I lần lượt là đầu tư của khu vực nhà nước, đầu tư từ khu vực tư nhân, đầu tư trực tiếp từ nước ngoài và tổng đầu tư.

Vì $I = GI + PI + FDI \Leftrightarrow GI + PI + FDI - I = 0$ nên mô hình này vi phạm ***Giả thiết 4*** do giữa các biến độc lập GI, DI, FDI, I có quan hệ đa cộng tuyến hoàn hảo.

Với ***Giả thiết 2*** được thỏa mãn thì từ mô hình (3.1) ta được

$$E(Y \mid X_2,...,X_k) = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Khi đó, β_1 còn được gọi là hệ số chặn và β_1 chính là giá trị trung bình của biến phụ thuộc Y khi các biến độc lập trong mô hình nhận giá trị bằng 0.

Các hệ số góc

$$\beta_j = \frac{\partial E\left(Y \mid X_2,...,X_k\right)}{\partial X_j}, j = \overline{2,k}$$

được gọi là các hệ số hồi quy riêng (partial coefficient), cho biết khi một biến $X_j, j = \overline{2,k}$ thay đổi một đơn vị, các biến khác cố định thì trung bình của Y thay đổi $\beta_j, j = \overline{2,k}$ đơn vị. Nếu có β_j nào đó bằng 0, ta nói biến Y không phụ thuộc vào biến độc lập X_j, có nghĩa là biến X_j không giải thích cho Y. Nếu tất cả β_j đều bằng 0, ta nói các biến độc lập đều không giải thích cho biến phụ thuộc Y, và hàm hồi quy trong trường hợp này được gọi là không phù hợp. Ngược lại, chỉ cần có ít nhất một biến độc lập giải thích cho biến phụ thuộc Y thì hàm hồi quy được gọi là phù hợp.

Ví dụ 2. Giả sử ta có mô hình hồi quy bội về lạm phát như sau

$$LP = 0.01 + 0.2M - 0.15GDP + U$$

trong đó LP, M, GDP lần lượt là tỷ lệ lạm phát, mức tăng trưởng cung tiền và mức tăng trưởng GDP (đơn vị %). Khi đó, ta có phiên giải từ mô hình trên như sau:

- Khi mức tăng trưởng cung tiền và GDP bằng 0 thì mức lạm phát trung bình là 0.01.
- Khi cung tiền **tăng (giảm)** 1% và mức tăng trưởng GDP không thay đổi thì lạm phát trung bình sẽ **tăng (giảm)** 0.2 đơn vị.
- Nếu GDP **tăng** 1% và cung tiền không thay đổi thì lạm phát trung bình sẽ **giảm** 0.15 đơn vị.

5.2. Phương pháp ước lượng OLS (Ordinary Least Squares)

Xét mô hình hồi quy tổng thể:

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + U \tag{5.2}$$

Để ước lượng các hệ số $\beta_1,...,\beta_k$ ta cần rút ra một mẫu ngẫu nhiên kích thước n từ tổng thể: $(X_{2i}, Y_{ki}), i = \overline{1,n}$. Khi đó, ta có

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots \beta_k X_{ki} + U_i$$

$$\widehat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$$

Ký hiệu phần dư (Residuals):

$$\widehat{U}_i = Y_i - \widehat{Y}_i$$

Chúng ta muốn xác định $\hat{\beta}_1,...,\hat{\beta}_k$ sao cho tổng bình phương các phần dư là bé nhất, có nghĩa là

$$f\left(\hat{\beta}_1,...,\hat{\beta}_k\right) = \sum_{i=1}^{n} \widehat{U}_i^{\,2} = \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ... - \hat{\beta}_k X_{ki}\right)^2 \to \min$$

Từ đó ta được

$$\hat{\beta} = \left(X^T X\right)^{-1}\left(X^T Y\right)$$

với

$$X = \begin{bmatrix} 1 & X_{21} & X_{31} & ... & X_{k1} \\ 1 & X_{22} & X_{32} & ... & X_{k2} \\ ... & ... & ... & ... & ... \\ 1 & X_{2n} & X_{3n} & ... & X_{kn} \end{bmatrix}, \; Y = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix}, \; \widehat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ ... \\ \hat{\beta}_k \end{bmatrix}$$

$$X^T X = \begin{bmatrix} n & \sum_{i=1}^{n} X_{2i} & \sum_{i=1}^{n} X_{3i} & ... & \sum_{i=1}^{n} X_{ki} \\ \sum_{i=1}^{n} X_{2i} & \sum_{i=1}^{n} X_{2i}^2 & \sum_{i=1}^{n} X_{2i}X_{3i} & ... & \sum_{i=1}^{n} X_{2i}X_{ki} \\ ... & ... & ... & ... & ... \\ \sum_{i=1}^{n} X_{ki} & \sum_{i=1}^{n} X_{ki}X_{2i} & \sum_{i=1}^{n} X_{ki}X_{3i} & ... & \sum_{i=1}^{n} X_{ki}^2 \end{bmatrix}, X^T Y = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{2i} Y_i \\ ... \\ \sum_{i=1}^{n} X_{ki} Y_i \end{bmatrix}$$

Ví dụ 3. Có số liệu quan sát của 15 cửa hàng khác nhau thuộc cùng một công ty kinh doanh cùng loại sản phẩm về lượng hàng bán được Y (tấn/tháng), chi phí quảng cáo X_2 (triệu đồng/tháng) và giá bán X_3 (ngàn đồng/kg) như sau:

Y_i	14	21	20	18	19	18	17	17	16	15	13	12	18,5	19	22
X_{2i}	5	9	8	7	8	8	6	6	5.7	5.5	4	3	7	8.2	9.5
X_{3i}	4	2.2	2.4	2.8	2.8	3	3.1	3.3	3.7	3.9	4.1	4.3	2.7	2.5	2

Khi đó

$$X^TX = \begin{bmatrix} 15 & \sum_{i=1}^{15} X_{2i} & \sum_{i=1}^{15} X_{3i} \\ \sum_{i=1}^{15} X_{2i} & \sum_{i=1}^{15} X_{2i}^2 & \sum_{i=1}^{15} X_{2i}X_{3i} \\ \sum_{i=1}^{15} X_{3i} & \sum_{i=1}^{15} X_{2i}X_{3i} & \sum_{i=1}^{15} X_{3i}^2 \end{bmatrix} = \begin{bmatrix} 15 & 99.9 & 46.8 \\ 99.9 & 713.23 & 293.64 \\ 46.8 & 293.64 & 153.52 \end{bmatrix}$$

$$\left(X^TX\right)^{-1} = \begin{bmatrix} 46.06 & -3.16 & -8.01 \\ -3.16 & 0.22 & 0.54 \\ -8.01 & 0.54 & 1.42 \end{bmatrix}, \; X^TY = \begin{bmatrix} 259.5 \\ 1801 \\ 780.65 \end{bmatrix} \Rightarrow \widehat{\beta} = \begin{bmatrix} 19.87 \\ 0.67 \\ -2.25 \end{bmatrix}$$

Vậy hàm hồi quy mẫu là

$$\widehat{Y}_i = 19.87 + 0.67X_{2i} - 2.25X_{3i}$$

Từ đó có nhận xét về ý nghĩa kinh tế của các tham số hồi quy:

- $\widehat{\beta_1} = 19.87$: Khi doanh nghiệp không quảng cáo và ngay cả khi bán giá cực thấp ($X_3 \to 0$) thì lượng hàng bán được tối thiểu trung bình khoảng 19.87 tấn/tháng.
- $\widehat{\beta_2} = 0.67 > 0$: Nếu giữ nguyên giá bán, khi **tăng (giảm)** mức quảng cáo một triệu đồng/tháng thì sản lượng tiêu thụ trung bình **tăng (giảm**) khoảng 0.67 tấn/tháng.
- $\widehat{\beta_3} = -2.25 < 0$: Nếu giữ nguyên chi phí quảng cáo, khi **tăng (giảm)** giá bán một ngàn đồng/kg thì sản lượng hàng bán được trung bình **giảm (tăng)** khoảng 2.25 tấn/tháng.

Sử dụng phần mềm Eviews, ta có bảng kết quả như sau:

Dependent Variable: Y
Method: Least Squares
Sample: 1 15
Included observations: 15

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	19.86862	2.694140	7.374756	0.0000
X2	0.669642	0.187374	3.573835	0.0038
X3	-2.252705	0.473382	-4.758751	0.0005
R-squared	0.983684	Mean dependent var	17.30000	

Adjusted R-squared	0.980965	S.D. dependent var	2.877251
S.E. of regression	0.396969	Akaike info criterion	1.166937
Sum squared resid	1.891008	Schwarz criterion	1.308547
Log likelihood	-5.752028	Hannan-Quinn criter.	1.165429
F-statistic	361.7403	Durbin-Watson stat	1.755206
Prob(F-statistic)	0.000000		

Để đo mức độ dao động và tương quan giữa các hệ số ước lượng được, sử dụng ma trận hiệp phương sai của hệ số hồi quy dạng tổng quát:

$$Cov\left(\widehat{\beta}\right)=\begin{bmatrix} \operatorname{var}\left(\widehat{\beta_1}\right) & Cov\left(\widehat{\beta_1},\widehat{\beta_2}\right) & ... & Cov\left(\widehat{\beta_1},\widehat{\beta_k}\right) \\ Cov\left(\widehat{\beta_2},\widehat{\beta_1}\right) & \operatorname{var}\left(\widehat{\beta_2}\right) & ... & Cov\left(\widehat{\beta_2},\widehat{\beta_k}\right) \\ ... & ... & ... & ... \\ Cov\left(\widehat{\beta_k},\widehat{\beta_1}\right) & Cov\left(\widehat{\beta_k},\widehat{\beta_2}\right) & ... & \operatorname{var}\left(\widehat{\beta_k}\right) \end{bmatrix}=\sigma^2\left(X^TX\right)^{-1}$$

Trong thực tế, thường không biết σ^2, do đó thay σ^2 bởi ước lượng điểm không chệch, tốt nhất $\widehat{\sigma^2}$ của nó trong các công thức ở trên:

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{n}\widehat{U}_i^2}{n-k}=\frac{RSS}{n-k}$$

Các sai số chuẩn của $\widehat{\beta_j}, j=\overline{1,k}$ là $Se\left(\widehat{\beta_j}\right)=\sqrt{Var\left(\widehat{\beta_j}\right)}$.

Từ số liệu của Ví dụ 3.3, dễ dàng tính được:

$$TSS=\sum_{i=1}^{n}Y_i^2-n\overline{Y}^2=11.5$$

$$ESS=\widehat{\beta}^T\left(X^TY\right)-n\overline{Y}^2=9.6$$

$$RSS=1.9$$

$$\hat{\sigma}^2=\frac{RSS}{n-k}=1.9/12=0.16$$

Ma trận hiệp phương sai của $\widehat{\beta}$ là:

$$Cov\left(\widehat{\beta}\right)=\begin{bmatrix} 7.27 & -0.5 & -1.26 \\ -0.5 & 0.04 & 0.08 \\ -1.26 & 0.08 & 0.22 \end{bmatrix}$$

Cũng có $\text{var}\left(\widehat{\beta_1}\right)=7.27, \text{var}\left(\widehat{\beta_2}\right)=0.04, \text{var}\left(\widehat{\beta_3}\right)=0.22$.

Sử dụng Eviews, ta được

	C	X2	X3
C	7.25838	-0.49728	-1.26153
X2	-0.49728	0.03510	0.08444
X3	-1.26153	0.08444	0.22409

Định lý Gauss-Markov: *Khi các giả thiết 1-4 được thỏa mãn thì các ước lượng thu được từ phương pháp OLS là các ước lượng tuyến tính, không chệch và có phương sai nhỏ nhất trong lớp các ước lượng tuyến tính không chệch (tính chất BLUE-Best Linear Unbiased Estimator).*

5.3. Độ phù hợp của hàm hồi quy mẫu

Tương tự như hồi quy hai biến, ta có định nghĩa các tổng bình phương độ lệch và hệ số xác định như sau:

Ký hiệu

$$TSS=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}Y_i^2-n\overline{Y}^2=Y^TY-n\overline{Y}^2$$

$$ESS=\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2=\hat{\beta}^T\left(X^TY\right)-n\overline{Y}^2$$

$$RSS=\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2=Y^TY-\hat{\beta}^T\left(X^TY\right)$$

Ta có $\quad TSS=ESS+RSS$

Hệ số xác định của hàm hồi quy bội cũng được định nghĩa là:

$$R^2=\frac{ESS}{TSS}=1-\frac{RSS}{TSS}=\frac{\hat{\beta}^T\left(X^TY\right)-n\overline{Y}^2}{Y^TY-n\overline{Y}^2}$$

Khi đó, $0\le R^2\le 1$. Chú ý rằng hệ số xác định trong mô hình hồi quy tuyến tính k biến là một hàm tăng theo số biến độc lập. Thật vậy, TSS không phụ thuộc vào số biến độc lập trong mô hình với bậc tự do là (n - 1), RSS là hàm giảm theo số biến độc lập trong mô hình với bậc tự do là (n - k), do đó khi số biến độc lập càng tăng thì hệ số xác định càng lớn và mô hình sẽ phức tạp hơn, khó phân tích hơn. Ngoài ra, khi số biến độc lập tăng lên sẽ làm tăng mối tương quan giữa các biến độc lập, đồng thời làm giảm bậc tự do của ESS, RSS. Do vậy cần cân nhắc cẩn thận trước khi thêm biến độc lập vào mô hình.

Với những chú ý trên, người ta điều chỉnh hệ số xác định bằng cách đưa thêm bậc tự do của các tổng bình phương vào công thức để được hệ số xác định hiệu chỉnh $\overline{R}^2$ (Adjustted R-Squared) sau đây:

$$\overline{R}^2 = 1 - \frac{\frac{RSS}{n-k}}{\frac{TSS}{n-1}} = 1 - \left(1 - R^2\right)\frac{n-1}{n-k} = R^2 + \left(1 - R^2\right)\frac{1-k}{n-k}$$

Người ta sử dụng $\overline{R}^2$ để xem xét có nên đưa thêm biến độc lập vào mô hình hay không. Việc đưa thêm biến độc lập vào mô hình là cần thiết khi trị số của $\overline{R}^2$ trong mô hình mới tăng lên, đồng thời hệ số hồi quy của biến độc lập đưa vào khác 0 và có ý nghĩa thống kê. Trở lại Ví dụ 3.3, sử dụng phần mềm Eviews để kiểm tra sự thay đổi của hệ số xác định trong hai trường hợp: (a) Lượng hàng bán được Y (tấn/tháng) chỉ phụ thuộc vào giá bán (ngàn đồng/kg) X_3 (triệu đồng/tháng), (b) Lượng hàng bán được Y (tấn/tháng) phụ thuộc vào cả chi phí quảng cáo X_2 (triệu đồng/tháng)và giá bán X_3 (ngàn đồng/kg):

Dependent Variable: Y
Method: Least Squares
Sample: 1 15
Included observations: 15

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X3	-3.863273	0.200042	-19.31231	0.0000
C	29.35341	0.639968	45.86703	0.0000

R-squared	0.966318	Mean dependent var	17.30000
Adjusted R-squared	0.963727	S.D. dependent var	2.877251
S.E. of regression	0.547984	Akaike info criterion	1.758423
Sum squared resid	3.903718	Schwarz criterion	1.852830
Log likelihood	-11.18817	Hannan-Quinn criter.	1.757417
F-statistic	372.9654	Durbin-Watson stat	1.411550
Prob(F-statistic)	0.000000		

Trường hợp (a)

Dependent Variable: Y
Method: Least Squares
Sample: 1 15
Included observations: 15

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	19.86862	2.694140	7.374756	0.0000
X2	0.669642	0.187374	3.573835	0.0038
X3	-2.252705	0.473382	-4.758751	0.0005

R-squared	0.983684	Mean dependent var	17.30000
Adjusted R-squared	0.980965	S.D. dependent var	2.877251
S.E. of regression	0.396969	Akaike info criterion	1.166937
Sum squared resid	1.891008	Schwarz criterion	1.308547
Log likelihood	-5.752028	Hannan-Quinn criter.	1.165429
F-statistic	361.7403	Durbin-Watson stat	1.755206
Prob(F-statistic)	0.000000		

Trường hợp (b)

Ta có $\overline{R}^2_{(a)} = 0.963727 < R^2_{(b)} = 0.980965$. Do đó, để xem việc đưa thêm biến X2 vào mô hình có phù hợp hay không, ta cần kiểm tra liệu có bác bỏ giả thuyết $H_0 : \beta_2 \neq 0$ hay không.

5.4. Một số dạng mô hình hồi quy nhiều biến

5.4.1. Mô hình logarit kép (log-log)

Xét mô hình hồi quy mũ (cho từng quan sát):

$$Y = \beta_1 X^{\beta_2} e^U \qquad (5.3)$$

Vì mô hình (3.3) không tuyến tính theo cả tham số và biến số nên ta không thể ước lượng nó theo phương pháp OLS. Để ước lượng được các tham số của mô hình này, ta cần chuyển về dạng tuyến tính theo tham số như sau:

Lấy logarit hai vế của (3.3), ta được

$$\ln Y = \ln \beta_1 + \beta_2 \ln X + U \qquad (5.4)$$

Đặt $\beta_1^* = \ln \beta_1$, (3.4) trở thành

$$\ln Y_i = \beta_1^* + \beta_2 \ln X_i + U_i,$$

gọi là mô hình log-log hay logarit kép. Đây là mô hình tuyến tính theo các tham số β_1^* và β_2 nên có thể ước lượng chúng bằng phương pháp OLS.

Nếu đặt $Y^* = \ln Y,\ X^* = \ln X$ thì

$$Y^* = \beta_1^* + \beta_2 X^* + U$$

Đặc điểm: Độ co giãn của Y đối với X được định nghĩa là

$$\varepsilon_{Y/X} = \frac{dY/Y}{dX/X} = \frac{dY}{dX} \times \frac{X}{Y}$$

Lấy vi phân hai vế của (3.4), khi đó

$$dY/Y = \beta_2 dX/X \Rightarrow \varepsilon_{Y/X} = \beta_2.$$

Như vậy, β_2 biểu thị hệ số co giãn của Y đối với X, do đó mô hình logarit kép còn được gọi là mô hình hệ số co giãn không đổi.

Một cách tổng quát, mô hình hồi quy dạng logarit kép có dạng như sau:

$$\ln Y = \beta_1 + \beta_2 \ln X_2 + \cdots + \beta_k \ln X_k + U$$

Với mỗi $j = 2,3,\ldots,k$, ta có $\beta_j = \dfrac{\frac{\partial Y}{Y}}{\frac{dX_j}{X_j}}$, có nghĩa là nếu X_j tăng (giảm) 1% và các yếu tố khác trong mô hình không đổi thì trung bình Y tăng (giảm) β_j%, và β_j được gọi là hệ số co giãn của Y theo X_j.

Ví dụ 4. Chẳng hạn có hàm cầu về thịt lợn như sau

$$\ln Q = 1.5 - 0.6 \ln P + U$$

Khi đó, hệ số co giãn về nhu cầu thịt lợn theo giá là -0.6, điều này có nghĩa là khi giá thịt lợn tăng 1% thì cầu trung bình về thịt lợn giảm 0.6%.

5.4.2. Mô hình bán logarit

Trong thực tế, có những trường hợp mô hình logarit kép không còn phù hợp, chẳng hạn quan hệ giữa tiền lương và số năm kinh nghiệm của người lao động, hoặc tiền lương và trình độ học vấn… Khi đó, người ta có thể sử dụng các mô hình bán logarit dưới đây:

Mô hình log-lin

Dạng mô hình:

$$\ln Y = \beta_1 + \beta_2 X + U \quad (5.5)$$

Trong mô hình này, hệ số β_2 được giải thích như sau: Khi X tăng 1 đơn vị thì Y trung bình tăng β_2%.

Trong nghiên cứu thực nghiệm, mô hình log-lin thích hợp với những tình huống như khảo sát tốc độ tăng trưởng hay giảm sút của các biến kinh tế tầm vĩ mô như: Dân số, lượng lao động, GDP, GNP, lượng cung tiền, năng suất, thâm hụt thương mại,…Chú ý thêm rằng mô hình log-lin chỉ thích hợp nếu số liệu chuỗi thời gian có tính chất dừng.

Ví dụ 5. Giả sử quan hệ giữa thu nhập (TN) và trình độ học vấn (Ed) là

$$\ln TN = 2.5 + 5.6Ed + U$$

Khi đó, ta có thể nói rằng cứ thêm mỗi năm đi học, mức thu nhập trung bình tăng 5.6%.

Ví dụ 6. Từ công thức tính lãi gộp

$$Y_t = Y_0(1+r)^t$$

Lấy logarit hai vế

$$\ln Y_t = \ln Y_0 + t\ln(1+r)$$

Đặt $\beta_1 = \ln Y_0, \beta_2 = \ln(1+r)$. Hơn nữa, nếu có thêm yếu tố ngẫu nhiên vào, ta được mô hình log-lin là

$$\ln Y_t = \beta_1 + \beta_2 t + U_t$$

Mô hình lin-log

Dạng mô hình:

$$Y = \beta_1 + \beta_2 \ln X + U \qquad 53.6)$$

Trong mô hình này, hệ số β_2 được giải thích như sau: Khi X tăng 1% thì Y trung bình tăng β_2 đơn vị.

Trong nghiên cứu thực nghiệm, ta có thể vận dụng mô hình này để khảo sát một số quan hệ như: Lượng cung tiền ảnh hưởng tới GDP, diện tích trồng trọt tác động tới sản lượng cây trồng, diện tích sử dụng của căn nhà tác động tới giá nhà…

Ví dụ 7. Có mô hình lin-log về quan hệ giữa số giờ mà người lao động muốn làm (L) và mức trả cho một giờ lao động (TL) như sau:

$$L = 7 + 0.6\ln TL + U$$

Kết quả trên cho thấy rằng khi mức trả cho một giờ lao động tăng 1% thì người lao động sẽ vui lòng làm thêm 0.6 giờ.

Nhận xét:

- Dạng hàm logarit thường được lựa chọn khi có gợi ý từ lý thuyết kinh tế về mối quan hệ giữa các biến số, chẳng hạn như dạng hàm Cobb-Douglas,…
- Dạng hàm logarit cũng thường được sử dụng khi các biến số đều nhận giá trị dương, hoặc biến số có phân phối đuôi lệch (thu nhập, mức lương, giá cổ phiếu,…). Việc lấy logarit giúp làm cho phân phối của sai số ngẫu nhiên gần với phân phối chuẩn, và do đó giúp tăng tính đối xứng của phân phối.
- Việc sử dụng mô hình dạng logarit có ưu thế là kết quả của ước lượng không phụ thuộc vào đơn vị đo của các biến số. Tuy nhiên, với những biến số mang cả giá trị âm (lợi nhuận công ty, lợi nhuận cổ phiếu) thì việc lấy logarit một cách trực tiếp là không thực hiện được.

5.4.5. Các ví dụ

Ví dụ 8. Để tìm hiểu giá bán thịt lợn (ngàn đồng) và giá bán thịt gà (ngàn đồng) có ảnh hưởng lên mức tiêu thụ thịt lợn hay không (kg/ngày), người ta khảo sát về mức tiêu thụ thịt lợn bình quân một ngày tại một siêu thị như sau:

Giá thị lợn	Giá thịt gà	Thịt lợn tiêu thụ
45	85	1200
49	81	1176
54	76	1152
58	73	1087
63	70	1045
69	72	1023
72	68	985
73	63	942
76	59	915
80	55	845
83	54	810
86	52	759

Ước lượng hàm hồi quy dạng:

+ Tuyến tính:

Dependent Variable: LUONGLONTIEUTHU
Method: Least Squares
Sample: 1 12
Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
GIATHITLON	-5.174643	1.989065	-2.601546	0.0287
GIATHITGA	6.969852	2.480846	2.809466	0.0204
C	874.0393	298.9491	2.923706	0.0169

R-squared	0.983099	Mean dependent var	994.9167
Adjusted R-squared	0.979344	S.D. dependent var	145.5125
S.E. of regression	20.91355	Akaike info criterion	9.130989
Sum squared resid	3936.388	Schwarz criterion	9.252216
Log likelihood	-51.78594	Hannan-Quinn	9.086107

		criter.	
F-statistic	261.7614	Durbin-Watson stat	0.999101
Prob(F-statistic)	0.000000		

+ **Mô hình lin-log:**

Dependent Variable: LUONGLONTIEUTHU
Method: Least Squares
Sample: 1 12
Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LOGX1	-213.3038	89.26039	-2.389680	0.0406
LOGX2	618.5673	116.0147	5.331799	0.0005
C	-707.8697	849.9657	-0.832822	0.4265

R-squared	0.984665	Mean dependent var	994.9167
Adjusted R-squared	0.981257	S.D. dependent var	145.5125
S.E. of regression	19.92154	Akaike info criterion	9.033798
Sum squared resid	3571.810	Schwarz criterion	9.155025
Log likelihood	-51.20279	Hannan-Quinn criter.	8.988916
F-statistic	288.9389	Durbin-Watson stat	1.227190
Prob(F-statistic)	0.000000		

+ **Mô hình log-log:**

Dependent Variable: LOGY
Method: Least Squares
Sample: 1 12
Included observations: 12

Variable	Coefficien t	Std. Error	t-Statistic	Prob.
LOGX1	- 0.111773	0.118087	-0.946533	0.3686
LOGX2	0.768074	0.153482	5.004325	0.0007

C	4.136796	1.124465	3.678903	0.0051

R-squared	0.974658	Mean dependent var	6.892549
Adjusted R-squared	0.969027	S.D. dependent var	0.149752
S.E. of regression	0.026355	Akaike info criterion	-4.221979
Sum squared resid	0.006251	Schwarz criterion	-4.100753
Log likelihood	28.33188	Hannan-Quinn criter.	-4.266862
F-statistic	173.0717	Durbin-Watson stat	1.072395
Prob(F-statistic)	0.000000		

Chương 6

ỨNG DỤNG MÔ HÌNH PHÂN TÍCH NHÂN TỐ KHÁM PHÁ

6.1. Phân tích mô tả và khám phá

6.1.1. Thống kê mô tả

Thống kê mô tả cung cấp các chỉ số cơ bản của biến số với dữ liệu của mẫu nghiên cứu. Hầu hết các nghiên cứu định lượng đều cần cung cấp các chỉ số thống kê mô tả để giúp người đọc hiểu về dữ liệu sử dụng. Các chỉ số và cách trình bày có thể khác nhau với biến định lượng và biến định danh.

Đối với các biến có giá trị liên tục (biến định lượng). Các nhà nghiên cứu thường cung cấp các chỉ số như giá trị trung bình, giá trị lớn nhất, giá trị nhỏ nhất và độ lêch chuẩn của biến. Trong một số trường hợp thì cần thêm giá trị trung vị.

Đối với các biến định danh. Các nhà nghiên cứu thường cung cấp các chỉ số như tần suất, tỷ lệ phần trăm trong tổng số, giá trị trung vị, giá trị yếu vị.

Ví dụ 1. Cho bảng thống kê mô tả của các biến định lượng

	Y	X2	X3
	Y	X2	X3
Mean	141.3333	20.41667	12.16667
Median	141.5000	20.50000	12.00000
Maximum	180.0000	26.00000	17.00000
Minimum	102.0000	15.00000	6.000000
Std. Dev.	23.20789	4.187825	3.352972
Skewness	-0.218444	-0.034924	-0.453123
Kurtosis	2.301560	1.401529	2.378514
Jarque-Bera	0.339345	1.279994	0.603763
Probability	0.843941	0.527294	0.739426
Sum	1696.000	245.0000	146.0000
Sum Sq. Dev.	5924.667	192.9167	123.6667
Observations	12	12	12

Bảng 6.1. Thống kê mô tả.

Trong đó

- Mean: Trung bình của lượng biến.
- Median: Trung vị của lượng biến.
- Maximun: Giá trị lớn nhất của lượng biến.

- Minimun: Giá trị nhỏ nhất của lượng biến.
- Std. Dev: Độ lệch chuẩn của lượng biến.
- Skewness: Hệ số bất đối xứng.
- Kurtosis: Hệ số nhọn.
- Jarque – Bera: Giá trị thống kê Jarque – Bera dùng để kiểm định phân phối chuẩn.
- Probability: Giá trị xác suất của thống kê Jarque – Bera dùng để kiểm định phân phối chuẩn.
- Sum: Tổng các giá trị của lượng biến.
- Sum Sq. Dev: Độ lệch chuẩn của tổng.
- Observations: Tổng số quan sát.

6.1.2. Ma trận hệ số tương quan

Các biến số có thể có tương quan với nhau, ma trận hệ số tương quan là một công cụ ban đầu để giúp các tác giả và người đọc quan sát về mối tương quan của từng cặp biến. Công đoạn này cũng giúp các tác giả nhận biết các hiện tượng bất thường hoặc đề phòng trường hợp đa cộng tuyến khi các biến độc lập có tương quan lớn. **Ví dụ 2.** Cho ma trận tương quan các biến định lượng như sau:

Correlation

	Y	X2	X3
	Y	X2	X3
Y	1.000000	0.782281	0.904627
X2	0.782281	1.000000	0.480173
X3	0.904627	0.480173	1.000000

Bảng 5.2. Ma trận tương quan giữa các biến.

Ý nghĩa. Ma trận tương quan cho biết mối tương quan giữa các biến trong mô hình. Ví dụ hệ số tương quan của X_2 và X_3 là 0,480173; hệ số tương quan của Y và X_2 là 0,782281; hệ số tương quan của Y và X_3 là 0,904627.

6.1.3. Phân tích nhân tố khám phá

Phân tích nhân tố khám khá là một phương pháp phân tích thống kê dùng để rút gọn một tập gồm nhiều biến quan sát có liên hệ với nhau thành một tập biến (gọi là các nhân tố) ít hơn để chúng có ý nghĩa hơn nhưng vẫn chứa đựng hầu hết nội dung thông tin của tập biến ban đầu (Hair và cộng sự, 1998).

Phân tích nhân tố khám phá thường được sử dụng nhiều nhất trong nghiên cứu khảo sát khi mà tác giả phải dùng nhiều câu hỏi để thu thập thông tin một vấn đề trừu tượng hơn, đặc biệt là những thông tin về tâm lý, thái độ, thậm chí hành vi. Kể cả khi chúng ta sử dụng thước đo đã được các tác giả trước phát triển và kiểm định thì cũng vẫn nên thực hiện kỹ thuật này xem liệu các mệnh đề/câu hỏi có "nhóm" đúng theo thước đo

ban đầu hay không. Kết quả phân tích nhân tố là cơ sở để tạo biến số cho các phân tích tiếp theo.

6.1.4. Phân tích độ tin cậy của thước đo

Phân tích độ tin cậy cho phép chúng ta xác định thuộc tính của thước đo mà chủ yếu là liệu các mệnh đề/câu hỏi của thước đo có "thống nhất" với nhau hay không. Thông thường các tác giả sử dụng chỉ số Cronbach' alpha từ 0,7 trở lên, song giá trị tối thiểu để thước đo có thể sử dụng được là 0,63 (DeVellis, 1990).

Phân tích độ tin cậy của thước đo thường được sử dụng cùng với phân tích nhân tố khám phá để quyết định các mệnh đề/câu hỏi cho từng thước đo. Lý tưởng nhất thước đo đủ cả ba điều kiện:

- Các câu hỏi/mệnh đề của thước đo được phát triển dựa trên lý thuyết hoặc đã được các tác giả trước xây dựng và kiểm định.
- Các câu hỏi/mệnh đề của thước đo "nhóm" cùng với nhau khi thực hiện phân tích nhân tố khám phá.
- Các câu hỏi/mệnh đề có chỉ số Crobach's alpha từ 0,7 trở lên, hoặc ít nhất cũng là 0,63.

6.2. So sánh nhóm

Một dạng nghiên cứu định lượng khá thông dụng là so sánh sự khác biệt giữa các nhóm về một hoặc một số chỉ số nào đó. Dưới đây là liệt kê các công cụ chính:

- T – test (kiểm định t) : được sử dụng để so sánh hai giá trị trung bình.
- ANOVA và ANCOVA (Analysis of Covariance) : Khi có nhiều hơn hai nhóm cần so sánh thì sử dụng ANOVA. Khi so sánh các nhóm, đồng thời kiểm soát tác động của một biến liên tục khác thì sử dụng ANCOVA.
- MANOVA (Multivariate Analysis of Variance): Tương tự như ANOVA nhưng được sử dụng khi có nhiều hơn một biến phụ thuộc và các biến phụ thuộc lại tương quan chặt với nhau.

Phụ lục 1. BẢNG CÂU HỎI KHẢO SÁT

1. Bảng câu hỏi khảo sát "200 hộ gia đình ở vùng nông thôn ĐBSCL năm 2016"

Xin kính chào quý ông (bà)

*Nghiên cứu "**Các nhân tố tác động đến thu nhập hộ gia đình**" nhằm đánh giá thực trạng thu nhập của hộ gia đình nông thôn, từ đó đề ra một số giải pháp để có thể nâng cao thu nhập hộ gia đình. Rất mong ông (bà) dành chút thời gian để trả lời phỏng vấn hoặc điền câu trả lời vào bảng khảo sát dưới đây. Những thông tin này sẽ được hoàn toàn giữ kín.*

Phần A. Thông tin tổng quát

Q1. Họ và tên chủ hộ:……………………………………………………………………….

Q2. Địa chỉ của hộ:………………………………………………………………………………

Q3. Trình độ văn hóa của chủ hộ:……………………………………………………….

Q4. Thành phần dân tộc của chủ hộ:………………………………………………………

Q5. Số nhân khẩu của hộ:………………………………………………………………..

Q6. Số người lao động chính của hộ trong 12 tháng qua:……………………………….

Q7. Xin ông (bà) cho biết thông tin của từng thành viên trong hộ:

Số thứ tự	Họ và tên	Quan hệ với chủ hộ	Giới tính	Tuổi	Trình độ học vấn	Nghề nghiệp chính	Kinh nghiệm	Hình thức nghề nghiệp chính	Nghề nghiệp phụ
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1		1							
2									
3									
4									
5									
6									
7									
8									
…									

Mã quy định các cột

(3) Quan hệ với chủ hộ: 1: Chủ hộ 2: Vợ/chồng 3: Con (gồm cả con đẻ, con nuôi) 4: Cha/mẹ (cả cha mẹ của vợ/chồng, cha mẹ nuôi nếu cùng ở trong một hộ) 5: Ông/bà 6: Cháu nội/ngoại 7: Dâu/rể 8: Anh/chị/em 9: Khác **(4) Giới tính:** 0: Nữ 1: Nam **(5) Tuổi:** *Ghi tuổi hiện tại* **(6) Trình độ học vấn:** 0: Không biết chữ 1: Cấp 1 2: Cấp 2 3: Cấp 3 4: Trung cấp 5: Cao đẳng 6: Đại học 7: Thạc sỹ 8: Tiến sỹ **(7) Nghề nghiệp chính:** 1: Nông nghiệp (Nông nghiệp, lâm nghiệp, thủy sản) 2: Công nghiệp và xây dựng 3: Dịch vụ (Thương nghiệp, vận tải, dịch vụ khác) **(8) Kinh nghiệp:** *Ghi số năm kinh nghiệm làm việc (nghề nghiệp chính)* **(9) Hình thức công việc:** 1: Tự làm cho gia đình 2: Làm thuê nhận tiền công, tiền lương. **(10) Nghề nghiệp phụ:** Việc làm chiếm thời gian nhiều thứ 2 trong 12 tháng qua 1: Tự làm cho gia đình 2: Làm thuê nhận tiền công, tiền lương.

Phần B. Thông tin về khả năng tiếp cận tín dụng

Q8. Trong 12 tháng qua, có ai trong hộ gia đình ông (bà) vay tiền tại các ngân hàng hay các tổ chức tín dụng không?

(Nếu có thì làm tiếp câu Q9, Q10, Q11; nếu không thì chuyển sang câu Q12).

☐ Có. ☐ Không.

Q9. Xin ông (bà) cho biết các khoản vay được vay từ nguồn nào?

☐ Ngân hàng nông nghiệp và PTNT ☐ Ngân hàng chính sách xã hội

☐ Ngân hàng khác, tổ chức tín dụng

☐ Các tổ chức chính trị xã hội

☐ Người cho vay cá thể

☐ Họ hàng, bạn bè

Q10. Giá trị khoản vay là bao nhiêu?...

Q11. Hộ gia đình ông (bà) vay các khoản tiền này để làm gì?

☐ Đầu tư cho sản xuất

☐ Xây, sửa chữa nhà

□ Tiêu dùng, sinh hoạt

□ Trả nợ

□ Chữa bệnh

□ Đi học

□ Khác (Xin vui lòng ghi rõ)…………………………………………………….

Phần C. Thông tin diện tích đất căn hộ

Q12. Gia đình ông (bà) có đất nông nghiệp, lâm nghiệp để trồng trọt, sản xuất không?
(Nếu có thì làm tiếp câu Q13, Q14; nếu không thì chuyển sang câu Q16).

□ Có. □ Không.

Q13. Xin ông (bà) cho biết diện tích thửa đất là boa nhiêu? Thuộc loại đất nào?

(1) Diện tích đất trồng cây hàng năm:…………………………………………………...

(2) Diện tích đất trồng cây lâu năm:…………………………………………………….

(3) Diện tích đất lâm nghiệp:……………………………………………………… (4) Diện tích đất dùng để chăn nuôi:……………………………………………………. (5) Diện tích đất nuôi trồng thủy sản:…………………………………………………...

(6) Tổng diện tích: (1)+(2)+(3)+(4)+(5):……………………………………………..

Q14. Gia đình ông (bà) có canh tác, sản xuất trên diện tích đất này không?
(Nếu có thì làm tiếp câu Q15; nếu không thì chuyển sang câu Q16).

□ Có. □ Không.

Q15. Nếu không sản xuất thì ông (bà) cho thuê hay bỏ hoang?

□ Cho thuê. □ Bỏ hoang.

Phần D. Thu nhập của gia đình

Q16. Tổng thu và tổng chi cho hoạt động sản xuất kinh doanh của hộ trong 12 tháng qua (*tính từ thời điểm đều tra về trước*)

Nguồn thu	**Tổng thu**	**Tổng chi**
1. Trồng trọt		
Cây hàng năm		
Cây lâu năm		
Sản phẩm phụ trồng trọt (thân, lá, ngọn cây, rơm rạ,…		
Sản phẩm trồng trọt khác (cây giống, cây cảnh,…)		
2. Chăn nuôi		
Gia súc		

Gia cầm		
Sản phẩm khác (trứng, con giống,..)		
Sản phẩm phụ chăn nuôi (phân, lông, da,…)		
3. Hoạt động dịch vụ nông nghiệp		
4. Lâm nghiệp và dịch vụ lâm nghiệp		
5. Thủy sản và dịch vụ thủy sản		
6. Các hoạt động sản xuất kinh doanh dịch vụ phi nông nghiệp		
7. Tiền lương, tiền công		
8. Các khoản khác (thu từ hái lượm, quà tặng, lãi tiết kiệm, tiền cho thuê nhà, lương hưu, trợ cấp,..		
Tổng cộng		

Q17. Theo ông (bà), Nhà nước cần hỗ trợ nào để giúp hộ gia đình phát triển kinh tế, nâng cao thu nhập?

☐ Phát triển cơ sở hạ tầng ☐ Vốn ☐ Kỹ thuật

☐ Đất sản xuất ☐ Tiêu thu sản phẩm, giá ổn định.

Xin chân thành cảm ơn quý ông (bà) đã tham gia trả lời câu hỏi. Chúc giá đình ông (bà) thật nhiều sức khỏa và hành phúc!

2. Bảng câu hỏi khảo sát "Ảnh hưởng của hình ảnh điểm đến tới lòng trung thành của khách du lịch: Nghiên cứu trường hợp điểm đến du lịch TP. HCM"

Chào Anh (Chị), Tôi là học viên cao học khóa 22 trường Đại học Kinh tế TP.HCM đang thực hiện đề tài luận văn cao học: "Ả**nh hưởng của hình ảnh điểm đến tới lòng trung thành của khách du lịch: Nghiên cứu trường hợp điểm đến du lịch TP. HCM"** *Trước tiên, xin cám ơn Anh (Chị) đã nhận lời tham gia cuộc phỏng vấn của chúng tôi. Cũng xin lưu ý mọi thông tin trung thực do Anh (Chị) cung cấp không có quan điểm nào là đúng hay sai và tất cả đều rất có giá trị cho nghiên cứu của chúng tôi, vì thế rất mong nhận được sự hợp tác của Anh (Chị).*

Tiếp theo, xin Anh (Chị) vui lòng cho biết mức độ đồng ý của Anh (Chị) về các phát biểu dưới đây với quy ước:

1	2	3	4	5
Rất không đồng ý	Không đồng ý	Trung lập	Đồng ý	Rất đồng ý

(Lưu ý: Mỗi phát biểu chỉ chọn 1 mức độ, Anh (Chị) đồng ý ở mức độ nào thì khoanh tròn vào mức độ tương ứng, trường hợp chọn nhầm xin Anh (Chị) gạch chéo X để hủy chọn và chọn lại).

Số TT	**Các tiêu thức**	**Mức độ đồng ý**				
		1	**2**	**3**	**4**	**5**
I. Môi trường cảnh quan						
1	TP. HCM có khí hậu, thời tiết dễ chịu quanh năm	1	2	3	4	5
2	TP. HCM có nhiều cảnh quan để tham quan, khám phá (sông, hồ, thác, biển, núi, vv)	1	2	3	4	5
3	Môi trường sống tại TP. HCM ít bị ô nhiễm khói bụi, tiếng ồn	1	2	3	4	5
4	TP. HCM có nhiều tòa nhà, công trình xây dựng hiện đại	1	2	3	4	5
5	TP. HCM có nhiều đường phố, đô thị sạch đẹp	1	2	3	4	5
II. Môi trường kinh tế, chính trị						
6	TP. HCM có tình hình chính trị ổn định	1	2	3	4	5
7	TP. HCM có các quy tắc, luật lệ rõ ràng	1	2	3	4	5
8	Tại TP. HCM ít xảy ra trộm cướp và tệ nạn xã hội	1	2	3	4	5
9	TP. HCM là trung tâm kinh tế phát triển, năng động	1	2	3	4	5
10	Anh/Chị cảm thấy TP. HCM an ninh và an toàn	1	2	3	4	5
III. Khả năng tiếp cận điểm đến						
11	Anh/Chị dễ dàng có được thông tin về điểm đến TP. HCM	1	2	3	4	5
12	Anh/Chị dễ dàng đến TP. HCM bằng các phương tiện vận chuyển khác nhau (ô tô, tàu thủy, tàu lửa, máy bay)	1	2	3	4	5
13	Dịch vụ đưa đón khách du lịch (xe ôm, taxi, xích lô) tại TP. HCM thuận tiện	1	2	3	4	5
14	Chuyến đi của anh/chị đến TP. HCM thuận lợi	1	2	3	4	5
15	Chi phí đi lại, vận chuyển hợp lý	1	2	3	4	5
IV. Lịch sử, văn hóa						
16	TP. HCM có các di tích lịch sử, di sản văn hóa đặc sắc	1	2	3	4	5

17	TP. HCM có các sự kiện, lễ hội văn hóa đa dạng và hấp dẫn	1	2	3	4	5
18	TP. HCM có nhiều bảo tàng lịch sử, văn hóa, nghệ thuật	1	2	3	4	5
19	Hoạt động của khách du lịch không bị hạn chế bởi rào cản ngôn ngữ	1	2	3	4	5
20	Người dân TP. HCM thân thiện và hiếu khách	1	2	3	4	5
V. Ẩm thực và mua sắm						
21	TP. HCM có nền ẩm thực phong phú, đa dạng	1	2	3	4	5
22	TP. HCM có nhiều món ăn độc đáo	1	2	3	4	5
23	Hoạt động mua sắm tại TP. HCM đa dạng, văn minh	1	2	3	4	5
24	Phong cách phục vụ tại TP. HCM chuyên nghiệp	1	2	3	4	5
25	Giá cả ăn uống, mua sắm phải chăng	1	2	3	4	5
VI. Cơ sở hạ tầng						
26	TP. HCM có hệ thống giao thông phát triển	1	2	3	4	5
27	Dịch vụ y tế, chăm sóc sức khỏe tại TP. HCM phát triển	1	2	3	4	5
28	Mạng lưới thông tin, viễn thông tại TP. HCM phát triển	1	2	3	4	5
29	Hệ thống khách sạn, nhà hàng TP. HCM đa dạng và có chất lượng	1	2	3	4	5
30	Hệ thống trung tâm lữ hành, công tu du lịch tại TP. HCM phát triển	1	2	3	4	5
VII. Giải trí						
31	TP. HCM có nhiều địa danh để thăm thú, tìm cảm giác lạ	1	2	3	4	5
32	TP. HCM có cuộc sống, hoạt động về đêm đa dạng (uống café, đi bar, coi kịch, nghe ca nhạc, xem phim, vv)	1	2	3	4	5
33	TP. HCM có các hoạt động ngoài trời, thể thao phong phú	1	2	3	4	5
34	TP. HCM có hoạt động vui chơi, tiêu khiển đặc sắc	1	2	3	4	5
VIII. Lòng trung thành của khách du lịch						
35	Anh/Chị ưu tiên lựa chọn điểm đến TP. HCM khi có cơ hội đi du lịch	1	2	3	4	5

36	Anh/Chị sẽ trở lại TP. HCM khi có cơ hội	1	2	3	4	5
37	Anh/Chị sẽ nỗ lực trở lại TP. HCM trong một ngày không xa	1	2	3	4	5
38	Anh/Chị sẽ giới thiệu người thân/ bạn bè đến du lịch tại TP. HCM	1	2	3	4	5

IX. THÔNG TIN CHUYẾN ĐI

39) Anh/Chị đã du lịch đến TP. HCM bao nhiêu lần (bao gồm cả chuyến đi này)?

☐ 1 lần
☐ 2-3 lần
☐ 4-5 lần
☐ Trên 5 lần

40) Lý do anh/chị đi du lịch TP. HCM? (có thể chọn nhiều câu trả lời)

☐ Tham quan
☐ Nghỉ dưỡng
☐ Vui chơi, giải trí
☐ Kinh doanh/ đi công tác
☐ Thăm người thân/ bạn bè
☐ Dự hội nghị/ triển lãm
☐ Đám cưới/ tuần trăng mật
☐ Mua sắm
☐ Tham dự sự kiện (thể thao, văn hóa, nghệ thuật, vv)
☐ Trên đường đến nơi khác
☐ Khác: (ghi rõ) …………………………………………………………………

41) Anh/Chị thường ở lại TP. HCM bao lâu?

☐ Dưới 1 ngày
☐ 1-2 ngày
☐ 3-5 ngày
☐ 1 tuần
☐ Trên 1 tuần

42) Anh/Chị du lịch đến TP. HCM bằng phương tiện nào? (có thể chọn nhiều câu trả lời)

☐ Xe máy
☐ Xe hơi
☐ Xe du lịch

☐ Taxi

☐ Tàu lửa

☐ Máy bay

☐ Khác: (ghi rõ) ……………………………………………………………

43) Anh/Chị đi du lịch TP. HCM với ai?

☐ Một mình

☐ Vợ/ chồng

☐ Gia đình và con cái

☐ Bạn bè/ họ hàng

☐ Đối tác kinh doanh

☐ Đòan du lịch

44) Anh/Chị biết đến TP. HCM thông qua kênh thông tin nào? (có thể chọn nhiều câu trả

lời)

☐ Internet

☐ Đại lý du lịch

☐ Ấn phẩm quảng cáo (brochure) / sách hướng dẫn du lịch (travel guidebook)/ tạp chí

☐ Bạn bè/ người thân

☐ Quảng cáo truyền hình

☐ Khác: (ghi rõ) ……………………………………………………………

45) Điểm du lịch nào tại TP. HCM mà anh/chị đã hoặc dự định đến?

(có thể chọn nhiều câu trả lời)

☐ Nhà hát lớn thành phố ☐ Khu du lịch Văn Thánh

☐ Thảo Cầm Viên ☐ Khu du lịch Bình Quới Thanh Đa

☐ Chợ Bến Thành ☐ Công viên văn hóa Đầm Sen

☐ Dinh Độc Lập ☐ Khu du lịch Suối Tiên

☐ Trụ sở UBND TP ☐ Khu giải trí BCR quận 9

☐ Nhà thờ Đức Bà ☐ Vườn Cò Thủ Đức

☐ Trung tâm mua sắm ☐ Cần Giờ

☐ Bảo tàng chứng tích chiến tranh ☐ Khu du lịch "Một thóang Việt Nam"

☐ Chùa bà Thiên Hậu ☐ Địa đạo Củ Chi

☐ Khác: (ghi rõ)……………………………………………………………

46) Điều gì tại TP. HCM mà anh/chị có ấn tượng tốt nhất?

…………………………………………………………………………………

…………………………………………………………………………………

X. THÔNG TIN CÁ NHÂN 47)

Giới tính:

☐ Nam ☐ Nữ 48) Tuổi:

☐ 18 - 35 ☐ 36 – 55

☐ Trên 55

49) Quốc tịch :

☐ Việt Nam ☐ Nước ngoài

50) Tình trạng hôn nhân:

☐ Chưa kết hôn ☐ Đã kết hôn 51) Trình độ:

☐ Chưa học qua cao đẳng - đại học

☐ Cao đẳng, đại học

☐ Sau đại học

52) Nghề nghiệp:

☐ Học sinh, sinh viên

☐ Công nhân, nhân viên văn phòng

☐ Cán bộ, viên chức, giảng viên

☐ Kinh doanh, buôn bán

☐ Khác: ………………...

53) Thu nhập hàng tháng:

☐ Dưới 5 triệu đồng

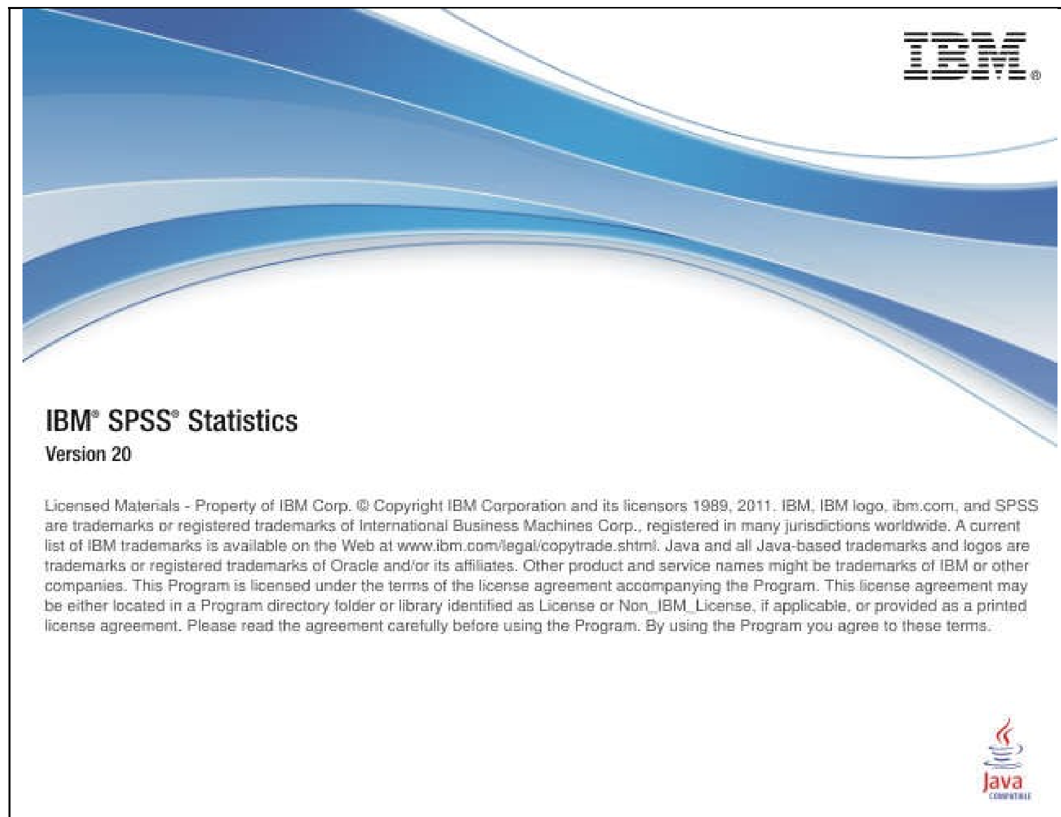
☐ Từ 5 triệu – dưới 10 triệu đồng

☐ Từ 10 triệu – dưới 20 triệu đồng

☐ Từ 20 triệu đồng trở lên

TRÂN TRỌNG CẢM ƠN SỰ HỖ TRỢ CỦA QUÝ ANH/CHỊ!

Phụ lục 2. HƯỚNG DẪN SỬ DỤNG PHẦN MỀM SPSS CƠ BẢN



1. Giới thiệu và cài đặt phần mềm SPSS

1.1. Giới thiệu phần mềm SPSS

SPSS (viết tắt của Statistical Package for the Social Sciences) là một chương trình máy tính phục vụ công tác thống kê. SPSS được sử dụng rộng rãi trong công tác thống kê xã hội.

Thế hệ đầu tiên của SPSS được đưa ra từ năm 1968. Thế hệ mới nhất là thế hệ 18 được giới thiệu từ tháng 8 năm 2008, có cả phiên bản cho các hệ điều hành Microsoft Windows, Mac, và Linux / UNIX.

SPSS là một hệ thống phần mềm thống kê toàn diện được thiết kế để thực hiện tất cả các bước trong các phân tích thống kê từ những tính toán trong thống kê mô tả (liệt kê dữ liệu, lập bảng tần số, lập biểu đồ, tính các đại lượng thống kê mô tả,...) đến thực hiện những bài toán trong thống kê suy diễn (kiểm định, tương quan, hồi quy, ...)

1.2. Cài đặt SPSS

Cho đến nay, SPSS có rất nhiều phiên bản khác nhau, phiên bản mới nhất là phiên bản 25.0. Với những tính toán thống kê phổ biến ta có thể cài SPSS với một trong các phiên bản 14.0, 16.0, 18.0 hay 20.0, ... Tất cả các phân tích trong tài liệu này được thực hiện trên phiên bản SPSS 20.0.

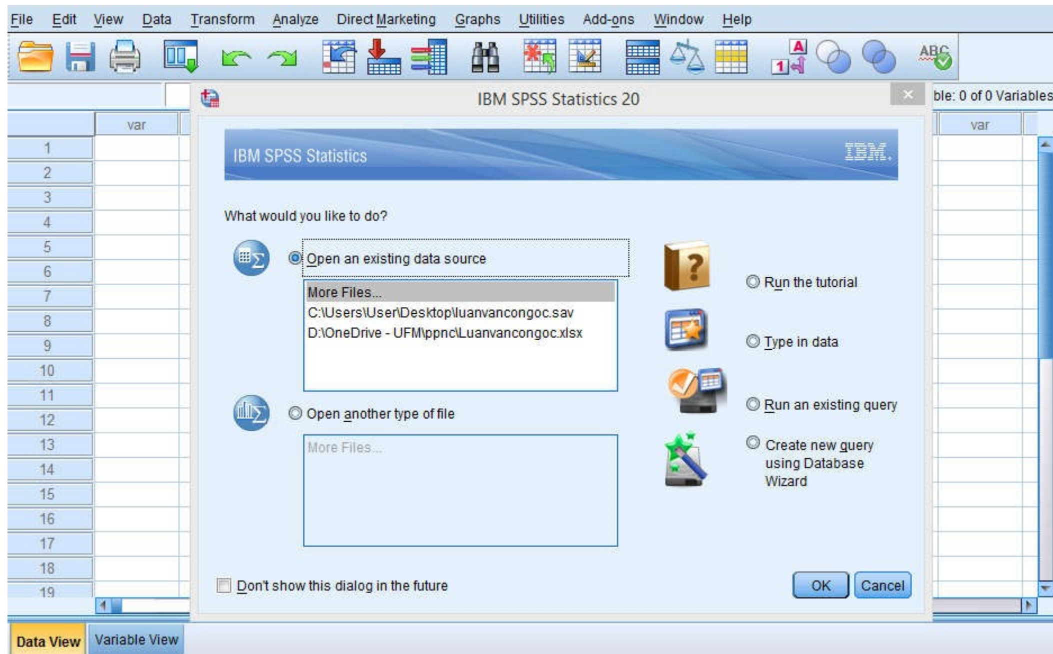
1.3. Khởi động SPSS

Để khởi động SPSS ta có thể thực hiện theo những cách sau:

+) Kích đúp vào biểu tượng SPSS trên màn hình;

+) Vào **Start□Program□ SPSS for Windows□SPSS 20.0 for Windows**

Sau khi khởi động, ta được giao diện sau:



Hình 1

2. Nhập dữ liệu trong SPSS

2.1. Nhập trực tiếp vào SPSSS

Để nhập trực tiếp một tập dữ liệu vào SPSS, ta thực hiện như sau:

Bước 1. Tại cửa sổ **IBM SPSS Statistics DaTa Editor**, ta ấn vào nút **Variable View** để khai báo thông tin về các thuộc tính của từng biến trong file dữ liệu. Trong một bảng Variable View

+) Các hàng là tên các biến;

+) Các cột là tên các thuộc tính của biến.

Các thuộc tính của một biến bao gồm:

+) Tên biến **(Name);** Loại dữ liệu **(Type)** +)

Số lượng con số hoặc chữ **(Width)**

+) Số lượng chữ số thập phân **(Decimals)**

+) Mô tả biến/nhãn biến **(Lable)** và nhãn trị số biến **(Values)**

+) Các giá trị khuyết thiếu do người sử dụng thiết lập **(Missing)**

+) Độ rộng của cột **(Columns)**

+) Căn lề **(Align)**

+) Thang đo của biến **(Measure)**

Khi khai báo hoặc chỉnh sửa các thuộc tính của biến trong cửa sổ **Variable View** cần chú ý một số điểm sau:

+) Tên biến phải bắt đầu bằng một chữ cái và không được kết thúc bằng một dấu chấm. Tên của biến là duy nhất, không được đặt trùng tên biến và tên biến không phân biệt chữ hoa, chữ thường.

+) Loại biến: Variable Type xác định loại dữ liệu đối với từng biến. Theo mặc định, mọi biến mới được giả sử là dạng số. Phụ thuộc vào loại dữ liệu được thu thập, ta có thể khai báo dưới những kiểu sau: dữ liệu là dạng số **(numeric)**, dấu phảy **(comma)**, dấu chấm **(dot)**, ghi chú khoa học **(Scientific notation)**, ngày tháng **(Date)**, đô-la **(Dollar)**, đơn vị tiền riêng **(custom currency)** và chuỗi **(string)**.

+) Nhãn của biến dùng để mô tả rõ hơn về tên của biến do tên của biến chỉ có độ dài tối đa là 8. Nhãn của biến có thể có độ dài đến 256 ký tự.

+) Ta có thể gán nhãn cho từng giá trị của biến. Tính năng này đặc biệt tiện lợi khi ta dùng các số để mã hóa các biến định tính. Ví dụ.... Biến được mã hóa như vậy có thể dùng cho nhiều phân tích khác nhau. Hơn nữa, với nhãn của các giá trị, kết quả tính ra sẽ được trình bày rõ ràng hơn.

+) Các giá trị khuyết thiếu do người sử dụng thiết lập **(Missing)**. Những giá trị không thích ứng với kiểu khai báo của biến sẽ được coi là giá trị khuyết thiếu. Đối với biến kiểu số, các ô trống được hiểu là giá trị khuyết và được đánh dấu bằng dấu phân cách thập phân. Nhiều thủ tục trong SPSS sẽ loại các giá trị khuyết ra khỏi các bước tính toán và các kết quả phân tích chỉ dựa trên phần số liệu không khuyết.

Ta có thể nhập đến 3 trị số khuyết riêng biệt, một phạm vi khoảng cách trị số khuyết hoặc một phạm vi cộng với một trị số khuyết riêng biệt;

Các phạm vi có thể được chỉ định cho các biến dạng số;

Các trị số khuyết cho các biến dạng chuỗi phải có độ dài không vượt quá 8 ký tự.

+) Số đo của biến có thể ở thang đo định danh **(Nominal)**, thang đo thứ bậc **(Ordinal)** hoặc thang đo khoảng, tỉ lệ **(gọi chung là Scale)**:

Thang đo định danh: Thang đo định danh dùng cho các biến định tính. Số đo của các biến này là các mã số để phân loại đối tượng. Giữa các mã số ở đây không có quan hệ hơn kém, chỉ dùng để đếm tần số xuất hiện của các biểu hiện. Một số ví dụ về thang đo này là: biến giới tính với các số đo là: Nam hoặc Nữ; biến màu sắc với số đo là: xanh, đỏ, tím, vàng,...; biến khu vực sống với các số đo: Thành phố, Thị xã, Nông thôn, Miền núi,...

Thang đo thứ bậc: Thang đo thứ bậc thường dùng cho các biến định tính, đôi khi dùng cho cả biến định lượng. Trong thang đo này giữa các số đo của các biến có quan hệ thứ bậc hơn kém. Tuy nhiên, sự chênh lệch giữa các số đo không nhất thiết bằng nhau. Ví dụ biến đánh giá thái độ đối với chất lượng dịch vụ mạng Internet tại nhà có số đo là: Không hài lòng, hài lòng, rất hài lòng.

Thang đo khoảng: Thang đo thứ bậc thường dùng cho các biến định lượng. Thang đo khoảng là thang đo thứ bậc có các khoảng cách đều nhau. Các phép tính cộng trừ đều có nghĩa nhưng không có giá trị không xác định một cách chính xác và không thể lấy tỉ lệ giữa các số đo. Ví dụ số đo nhiệt độ, số đo chỉ số IQ, chỉ số EQ,...

Thang đo tỉ lệ: Thang đo tỉ lệ dùng cho các biến định lượng. Thang đo tỉ lệ là thang đo khoảng, hơn nữa thang đo này có giá trị không xác định một cách chính xác và có thể lấy tỉ lệ giữa các số đo. Ví dụ về thang đo này đơn vị đo tiền tệ (VND, dollar, pound, yen,...); đơn vị đo chiều dài (cm, m, km,...); đơn vị đo khối lượng (kg, tấn, tạ, yến,...).

Bước 2. Tại cửa sổ **Data Editor**, nhấn vào nút **Data View** để nhập từng dữ liệu trong mỗi ô. Trong bảng Data View +) Mỗi cột là mỗi biến **Variable**;

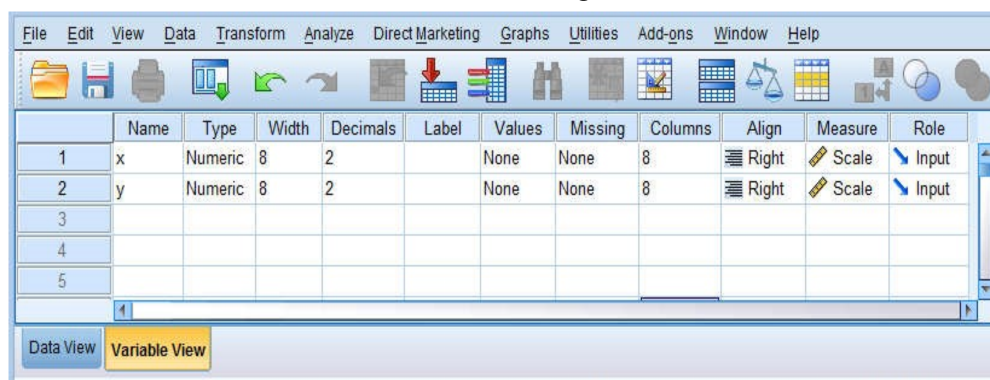
+) Mỗi hàng là một đối tượng **cases**.

Ví dụ 1. Cho dữ liệu

X	6	10	12	14	16	18	22	24	26	32
Y	40	44	46	48	52	58	60	68	74	80

Bảng 1

Sau khi đã khai báo các thuộc tính của các biến trong cửa sổ Variable View



Hình 2

Ta vào cửa sổ Data View nhập giá trị cho từng biến như đã được điều tra.

	x	y	var	var	var	var
1	6.00	40.00				
2	10.00	44.00				
3	12.00	46.00				
4	14.00	48.00				
5	16.00	52.00				
6	18.00	58.00				
7	22.00	60.00				
8	24.00	68.00				
9	26.00	74.00				
10	32.00	80.00				

Hình 3

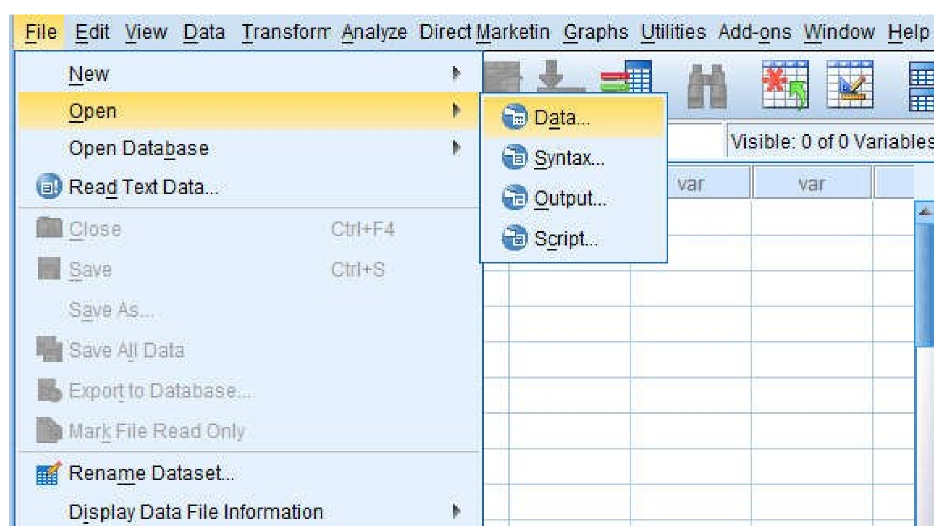
Để lưu dữ liệu vừa được tạo ra dưới đuôi của **SPSS.sav**, ta vào **File□Save as** để đánh tên file cần lưu và thư mục để lưu trong máy tính.

2.2. Đọc dữ liệu từ những file có sẵn

SPSS có thể đọc dữ liệu từ rất nhiều kiểu file khác nhau. Ngoài dữ liệu dưới dạng file.sav của SPSS, SPSS còn cho đọc file dữ liệu dưới một số dạng thông dụng như **file.xls (Excel File)**, **file.txt (Text File)** và file dữ liệu tạo ra từ một số phần mềm thống kê khác như **file.dta (Stata File), file.wf1 (Eviews Workfile), ...**

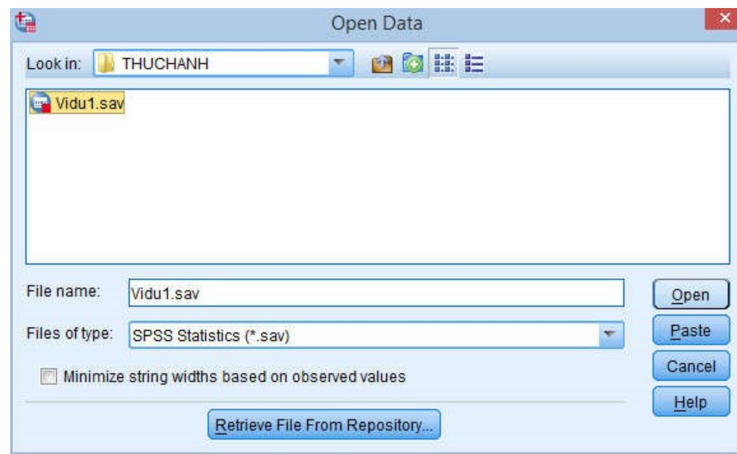
2.2.1. Đọc dữ liệu từ file.sav

Để đọc dữ liệu từ một file.sav, chẳng hạn file dữ liệu **Vidu1.sav**, ta vào **File□ Open□Data** và chọn đến thư mục để file dữ liệu và mở file:



Hình 4

Kích đúp Data…



Hình 5

Chọn **Vidu1** rồi kích Open. Ta được

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Visible: 2 of 2 Variables

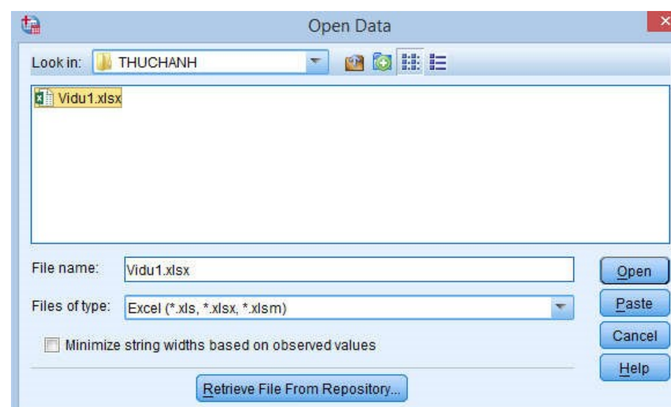
	x	y	var	var	var	var	var
1	6.00	40.00					
2	10.00	44.00					
3	12.00	46.00					
4	14.00	48.00					
5	16.00	52.00					
6	18.00	58.00					
7	22.00	60.00					
8	24.00	68.00					
9	26.00	74.00					
10	32.00	80.00					
11							

Data View | Variable View

Hình 6

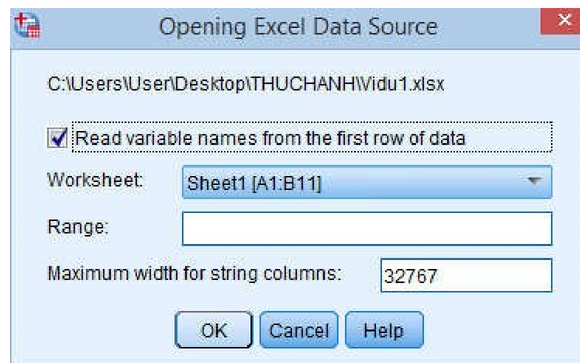
2.2.2. Đọc dữ liệu từ file .xls

Để đọc dữ liệu từ một file.xls, chẳng hạn file dữ liệu **Vidu1.xlsx**, ta vào **File Open Data** và chọn đến thư mục để file dữ liệu và mở file:



Hình 7

Chọn **Vidu1** rồi kích Open màn hình xuất hiện cửa sổ sau



Hình 8

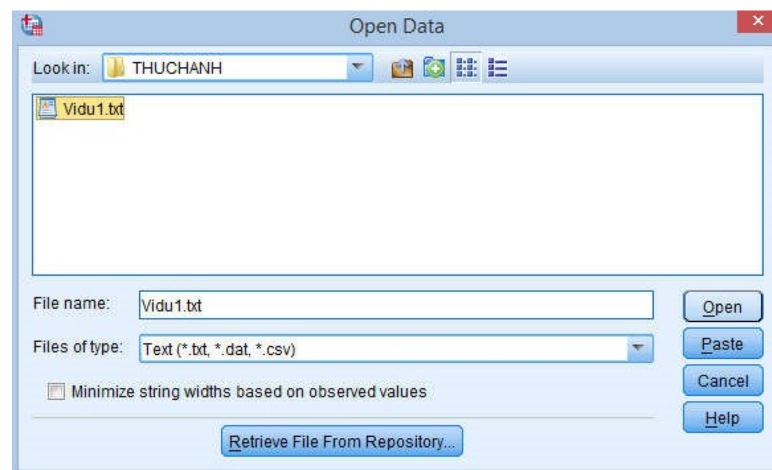
Kích **OK**, ta được

	x	y
1	6.00	40.00
2	10.00	44.00
3	12.00	46.00
4	14.00	48.00
5	16.00	52.00
6	18.00	58.00
7	22.00	60.00
8	24.00	68.00
9	26.00	74.00
10	32.00	80.00

Hình 9

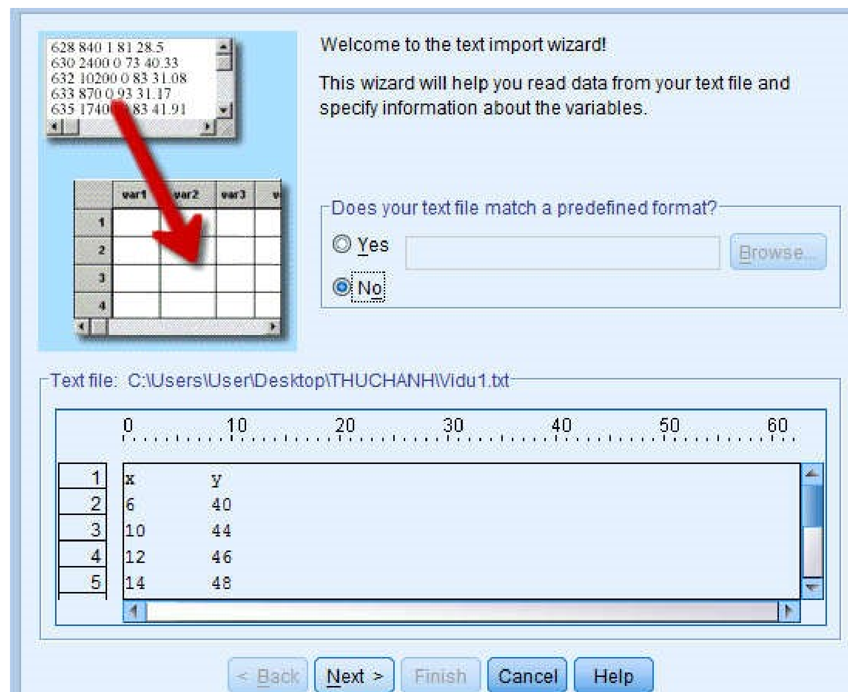
2.2.3. Đọc dữ liệu từ file.txt

Để đọc dữ liệu từ một file.txt, chẳng hạn file dữ liệu **Vidu1.txt**, ta vào **File □Open □Data** và chọn đến thư mục để file dữ liệu và mở file:



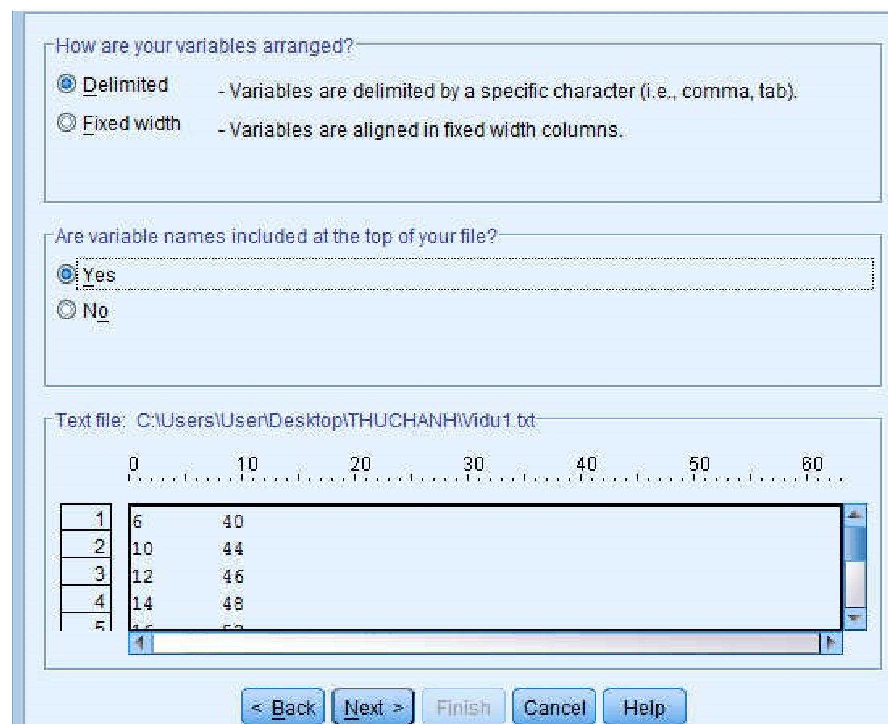
Hình 10

Khi màn hình trên hiện ra hộp thoại sau ta ấn vào **Open**



Hình 11

Ta ấn **Next**, cửa sổ sau xuất hiện



Hình 12

Khi trên màn hình hiện ra hộp thoại sau ta chọn **Yes** cho câu hỏi **Are variable names include at the top of your file?** và sau đó ấn vào **Next**. Tiếp tục nhấn **Next□ Next□Next□Finish**, ta được:

	x	y
1	6.00	40.00
2	10.00	44.00
3	12.00	46.00
4	14.00	48.00
5	16.00	52.00
6	18.00	58.00
7	22.00	60.00
8	24.00	68.00
9	26.00	74.00
10	32.00	80.00

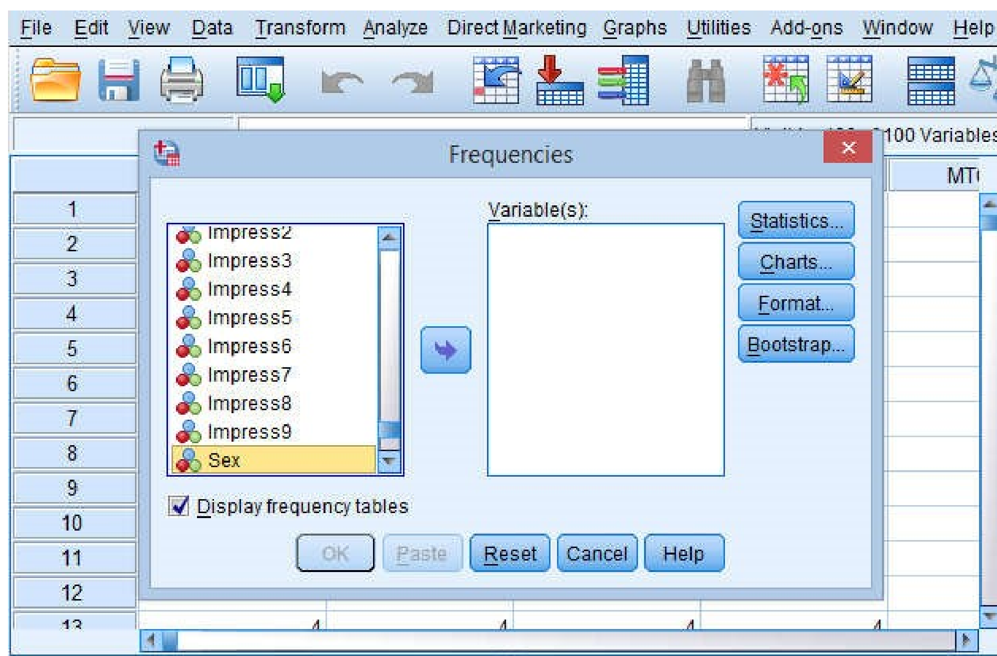
Hình 13

3. Lập bảng tần số và tính các đại lượng thống kê mô tả

3.1. Lập bảng tần số

Để lập bảng tần số cho các biến định tính hoặc dữ liệu định lượng có ít biểu hiện trong tập dữ liệu, chẳng hạn Ví dụ 1.

Bước 1: Vào **Analyze□Descriptive Statistics □ Frequencies...**. Màn hình hiện ra hộp thoại:



Hình 14

Bước 2: Chọn các biến cần lập bảng tần số bên trái đưa vào khung **Variable** bên phải và ấn **OK**.

Ta có thể tiến hành lập bảng tần số cho một biến hoặc một số biến cùng một lúc, chẳng hạn khi chọn biến Sex trong cửa sổ **Output** nhận được kết quả sau: **Sex**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	172	47.8	47.8	47.8
	2	188	52.2	52.2	100.0
	Total	360	100.0	100.0	

Bảng 2

3.2. Tính các đại lượng thống kê mô tả

Trong một tập dữ liệu ta có thể tính các đại lượng thống kê mô tả của tập dữ liệu như +)

Các đại lượng đo độ tập trung: trung bình cộng, trung vị, mode.

+) Các đại lượng đo độ phân bố: tứ phân vị và phân vị thứ p.

+) Các đại lượng đo độ phân tán: khoảng biến thiên, độ trải giữa, phương sai, độ lệch chuẩn.

+) Các đại lượng mô tả hình dáng của tập dữ liệu: hệ số bất đối xứng Skewness, hệ số đo đọ nhọn Kurtosis.

Trước hết ta nhắc lại định nghĩa và công thức tính toán những đại lượng này.

+) Trung bình cộng đơn giản được tính bằng cách cộng tất cả các giá trị quan sát của tập dữ liệu rồi chia cho số quan sát của tập dữ liệu đó.

+) Trung vị là giá trị đứng giữa của tập dữ liệu đã được sắp thứ tự. Như vậy, không kể trung vị sẽ có 50% số quan sát của tập dữ liệu có giá trị lớn hơn trung vị và 50% số quan sát của tập dữ liệu có giá trị nhỏ hơn trung vị.

+) Mode của một tập dữ liệu là giá trị xuất hiện nhiều nhất trong tập dữ liệu.

+) Tứ phân vị chia tập dữ liệu đã sắp xếp theo trật tự tăng dần thành bốn phần có số quan sát bằng nhau. Tứ phân vị bao gồm ba tứ phân vị: tứ phân vị thứ nhất Q_1, tứ phân vị thứ hai Q_2, (chính là trung vị) và tứ phân vị thứ ba Q_3. Trong tập dữ liệu có 25% số quan sát của tập dữ liệu nhỏ hơn hoặc bằng Q_1, 50% số quan sát của tập dữ liệu nhỏ hơn hoặc bằng Q_2 và có 75% số quan sát của tập dữ liệu nhỏ hơn hoặc bằng Q_3.

+) Phân vị thứ p của một tập dữ liệu đã được sắp thứ tự là giá trị chia tập dữ liệu thành hai phần, một phần gồm p% số quan sát có giá trị nhỏ hơn hoặc bằng phân vị thứ p, phần còn lại có (100 p)% số quan sát lớn hơn hoặc bằng phân vị thứ p.

+) Khoảng biến thiên của một tập dữ liệu là hiệu giữa giá trị lớn nhất và giá trị nhỏ nhất của tập dữ liệu.

+) Độ trải giữa của một tập dữ liệu là hiệu độ chênh lệch giữa tứ phân vị thứ ba và tứ phân vị thứ nhất của tập dữ liệu.

+) Phương sai của một tập dữ liệu tổng thể, ký hiệu là σ^2, được xác định bởi công thức:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2,$$

trong đó μ là trung bình của tổng thể và N là số quan sát trong tổng thể.

+) Phương sai của một tập dữ liệu mẫu, ký hiệu là s^2, được xác định bởi công thức:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

trong đó $\bar{x}$ là trung bình của mẫu và n là số quan sát trong mẫu.

+) Độ lệch chuẩn của một tập dữ liệu tổng thể, kí hiệu là σ, là căn bậc hai của phương sai của tổng thể:

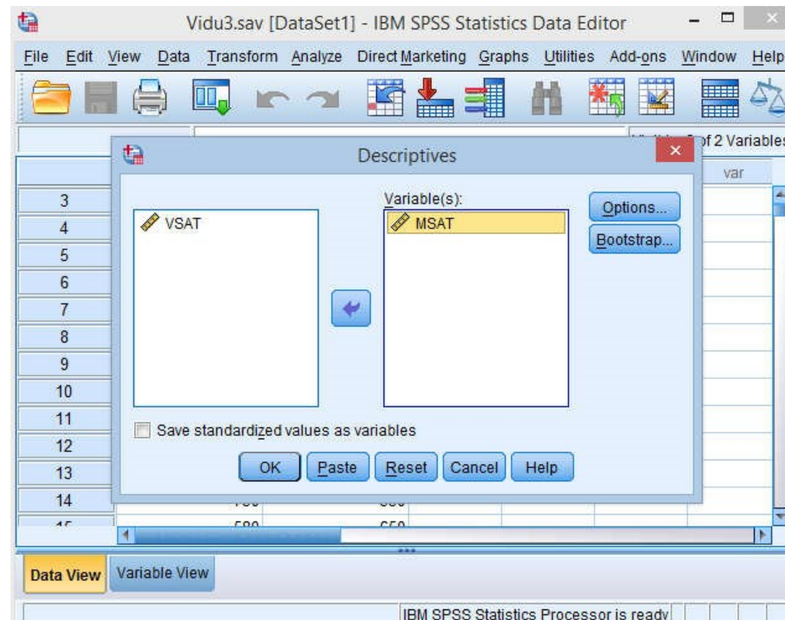
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}.$$

+) Độ lệch chuẩn của một tập dữ liệu mẫu, kí hiệu là s, là căn bậc hai của phương sai mẫu:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

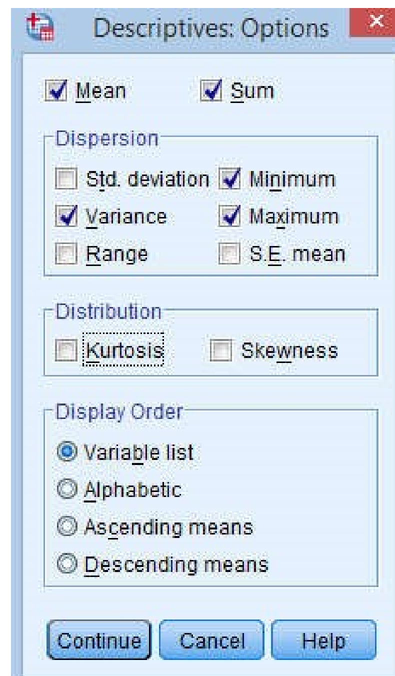
Chẳng hạn, để tính những đại lượng thống kê mô tả như: trung bình, trung vị, phương sai, độ lệch chuẩn của các biến từ file Vidu3.sav.

Bước 1. Vào **Analyze→Descriptive Statistics→Descriptive...** và khi màn hình sẽ hiện ra hộp thoại sau thì ta chọn biến MSAT để đưa vào khung **Variable(s)** ở bên phải như sau:



Hình 15

Bước 2. Ấn tiếp vào nút **Options** để chọn tính những đại lượng thống kê mô tả cho biến MSAT và màn hình hiện ra hộp thoại:



Hình 16

Trong đó

Mean	Trung bình cộng
Sum	Tổng các giá trị của biến
Std.Deviation	Độ lệch chuẩn

Minimum	Giá trị nhỏ nhất
Maximum	Giá trị lớn nhất
Variance	Phương sai
Range	Khoảng biến thiên
SE mean	Độ lệch chuẩn của trung bình mẫu
Kurtosis	Hệ số đo độ nhọn
Skewness	Hệ số bất đối xứng

Bảng 3

Ta muốn tính đại lượng nào thì chọn vào đại lượng đó để hiện kết quả ra cửa sổ Output sau đó ấn **Continue** và khi màn hình hiện ra hội thoại cũ thì ấn **OK.** Chẳng hạn, ta chọn như hình cửa sổ Output sẽ cho ta kết quả sau:

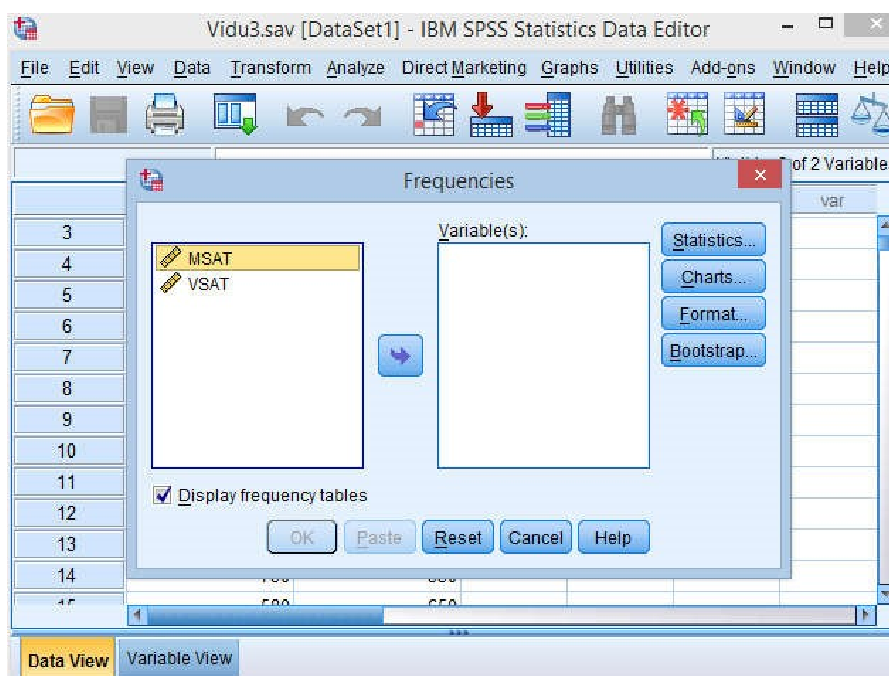
Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean	Variance
MSAT	427	330	770	241820	566.32	8671.191
Valid N (listwise)	427					

Bảng 4

Tính tứ phân vị hay phân vị thứ p tùy ý của tập dữ liệu, chẳng hạn cho biến MSAT:

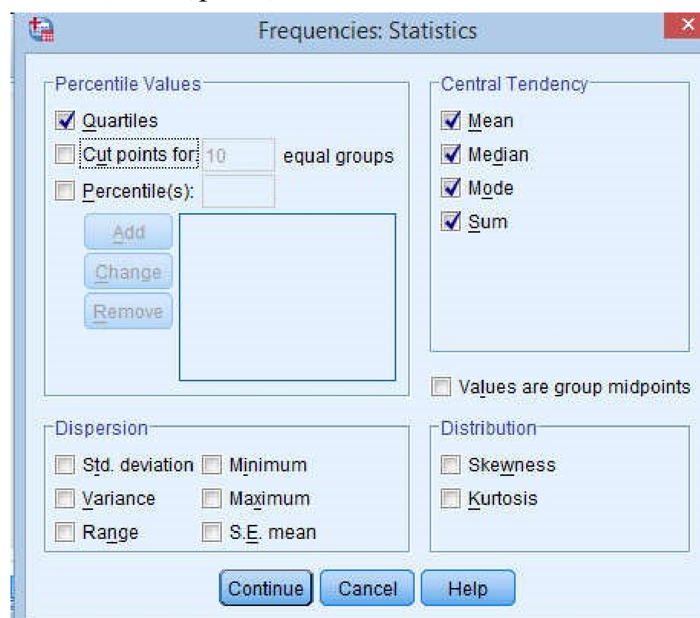
Bước 1. Vào **Analyze□Descriptive Statistics□Frequencies...** và màn hình hiện ra hộp thoại:



Hình 17

Bước 2. Chọn các biến định lượng cần tính bên trái đưa vào khung **Variable** bên phải, ấn **Statistics** và nếu ta không muốn đưa bảng tần số thì không chọn ô **Display frequency tables**.

Bước 3. Khi màn hình hiện ra hộp thoại:



Hình 18

ta chọn những đại lượng cần tính, ấn **Continue** trở về hộp thoại trước ấn **OK**. Ta được

Statistics

MSAT

N	Valid	427
	Missing	0
Mean		566.32
Median		570.00
Mode		530
Sum		241820
	25	510.00
Percentiles	50	570.00
	75	640.00

Bảng 5 Kết quả tính toán này trên SPSS cho ta thấy:

+) Trung vị của MSAT là 570, mode của tuổi là 530;

+) Ba tứ phân vị là: Q 510□ ; Q_2 □ 570; Q_3 □ 640.;

4. Biểu đồ

4.1. Một số loại biểu đồ

+) Biểu đồ thanh, biểu đồ tròn: áp dụng cho dữ liệu định tính

+) Biểu đồ hộp và râu, biểu đồ tần số, biểu đồ tán xạ, biểu đồ gấp khúc,...: áp dụng cho dữ liệu định lượng.

4.2. Biểu đồ thanh

+) Biểu đồ thanh được dùng cho dữ liệu định danh hay thứ bậc.

+) Chiều cao của mỗi thanh biểu diễn tần số hay tần suất của trường hợp biểu diễn bởi thanh đó.

+) Có thể áp dụng để biểu diễn cho một hay nhiều tập dữ liệu trên cùng 1 biểu đồ.

Cách vẽ biểu đồ thanh

Bước 1. Vào menu **Graph□Bar** để mở cửa sổ **Bar Chart**

Bước 2. Tại cửa sổ **Bar Chart**

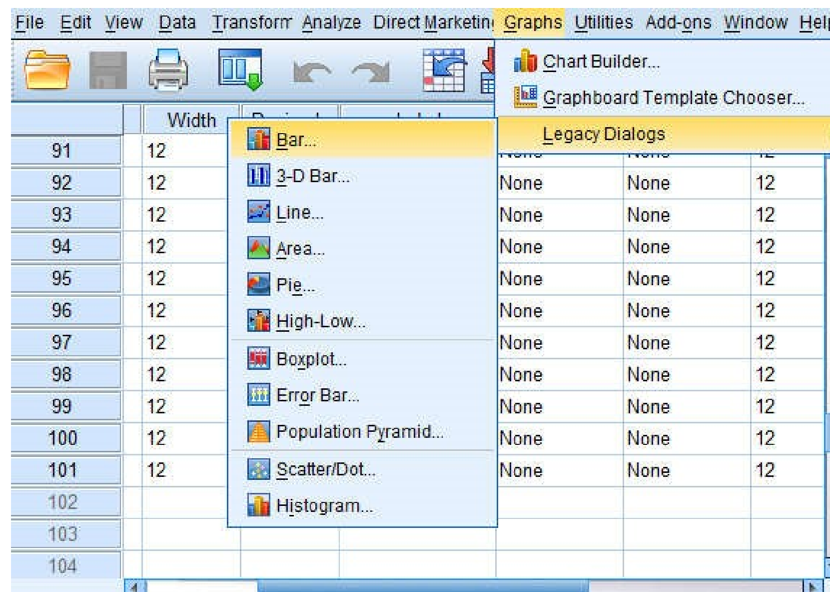
+) Chọn **Simple**: nếu biểu diễn tập dữ liệu một biến

+) Chọn **Clustered**: nếu muốn biểu diễn tập dữ liệu của một biến được chia nhóm bởi một biến khác, các thanh đứng kề nhau

+) Chọn **Stacked**: nếu muốn biểu diễn tập dữ liệu của một biến được chia nhóm bởi một biến khác, các thanh chồng lên nhau

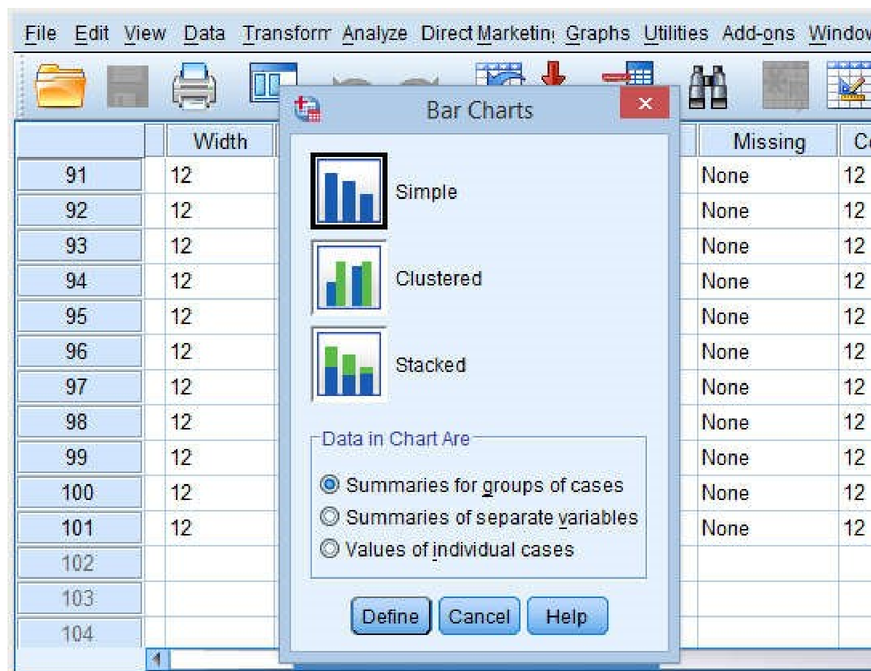
Vẽ biểu đồ thanh biểu diễn một biến: Để vẽ biểu đồ thanh biểu diễn biến nghề nghiệp (Career) trong file **Vidu2.sav** ta tiến hành như sau

Bước 1. Từ cửa sổ **Data View□ Graphs□ Legacy Dialogs□ Bar**



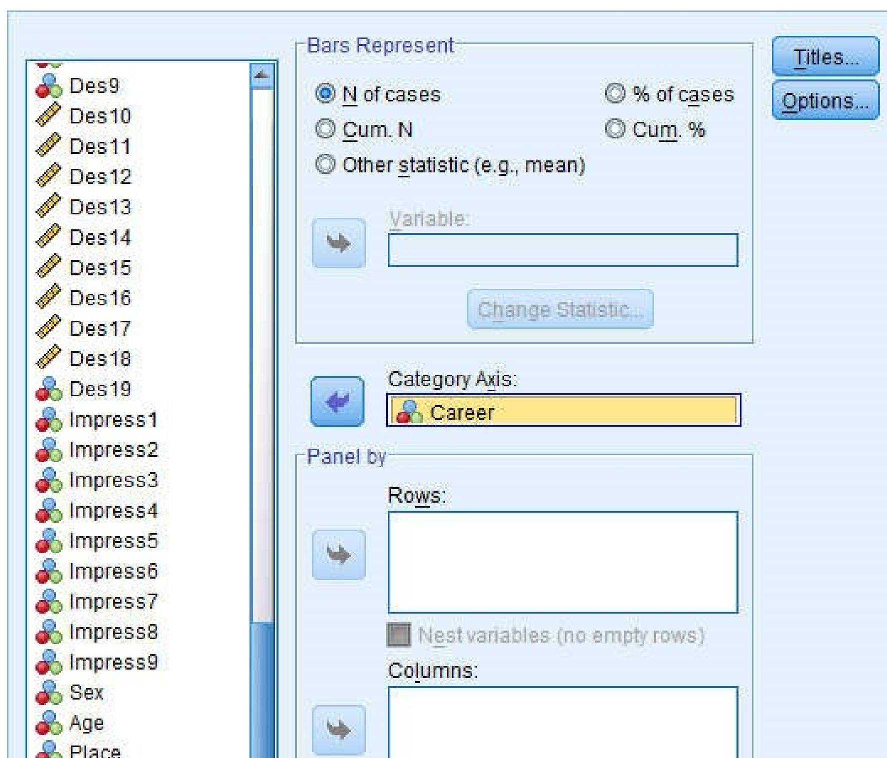
Hình 19

Kích đúp Bar, ta được



Hình 20

Bước 2. Tại hộp thoại **Bar Charts**, chọn **Simple**. Tại khu vực **Data in Chart Area** chọn **Summary for group of cases**. Ấn **Define**. Xuất hiện cửa sổ



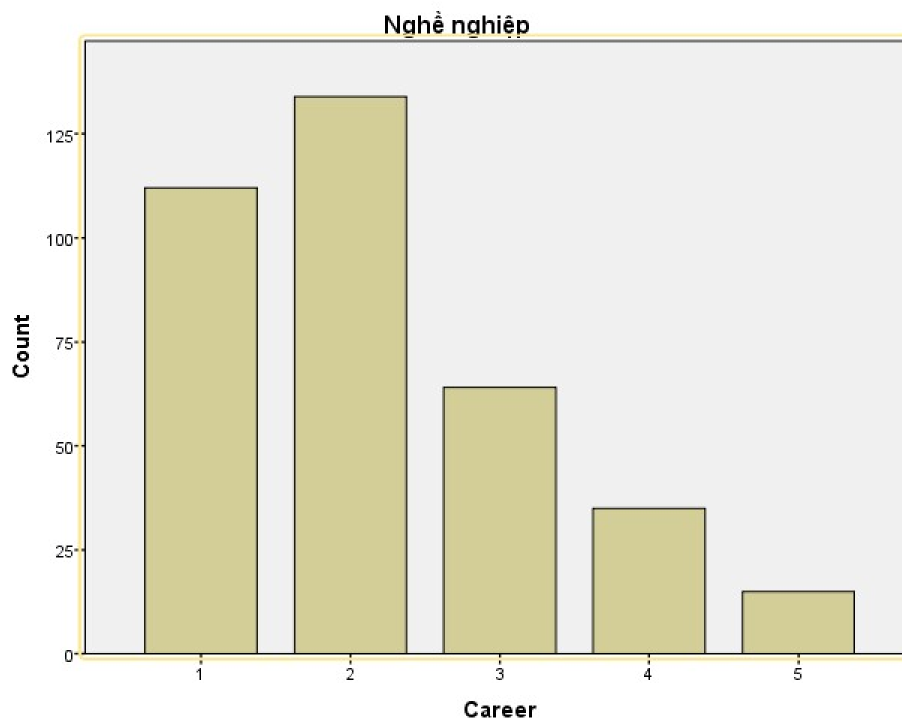
Hình 21

Bước 3. Đưa biến Career vào khung **Category Axis.** Lựa chọn **N of cases** để trục tung biểu diễn tần số, % **of cases** để trục tung biểu diễn tần suất.

Bước 4. Ấn nút **Title** để đặt tiêu đề cho hình rồi ấn **Continue.**

Hình 22

Bước 5. Ấn **OK**. Ta được

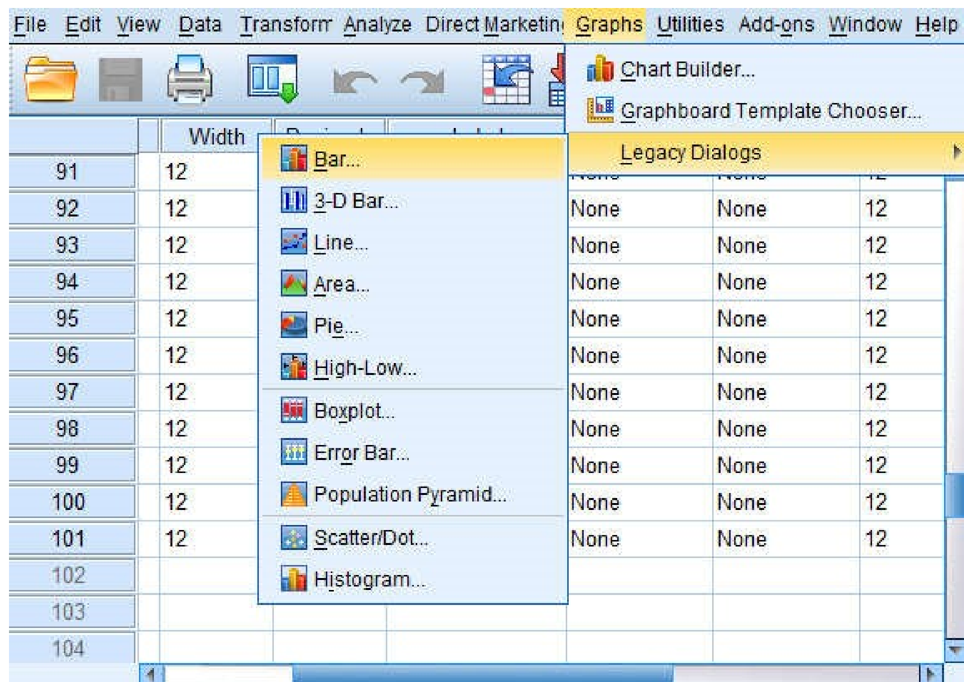


Hình 23

Ví dụ: vẽ biểu đồ thanh biểu diễn một biến được phân tách bởi một biến khác

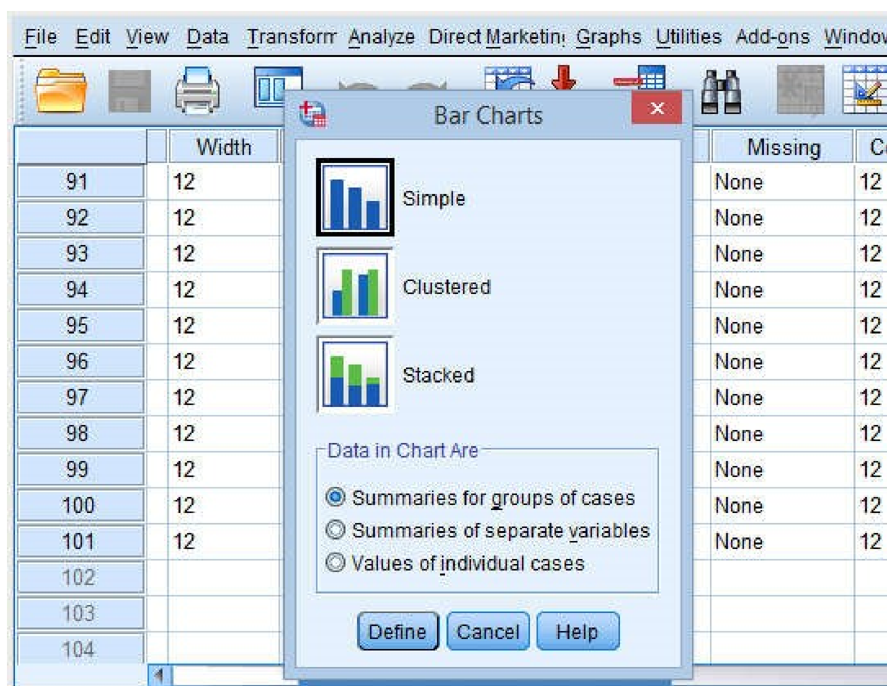
Để vẽ biểu đồ thanh biểu diễn biến nghề nghiệp (Career) theo từng nhóm nam, nữ (biến nghề nghiệp được phân tách bởi biến giới tính) trong file **Vidu2.sav** ta tiến hành như sau:

Bước 1. Từ cửa sổ **Data View Graphs Legacy Dialogs Bar**



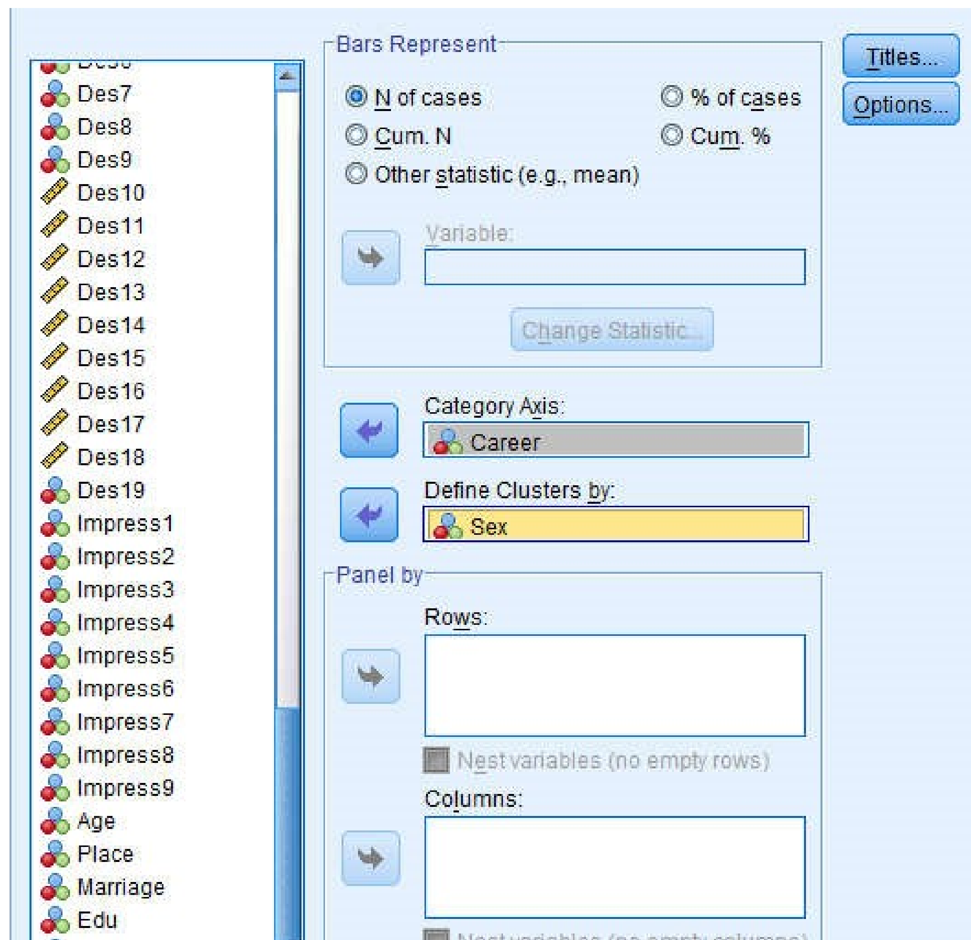
Hình 24

Kích đúp Bar, ta được



Hình 25

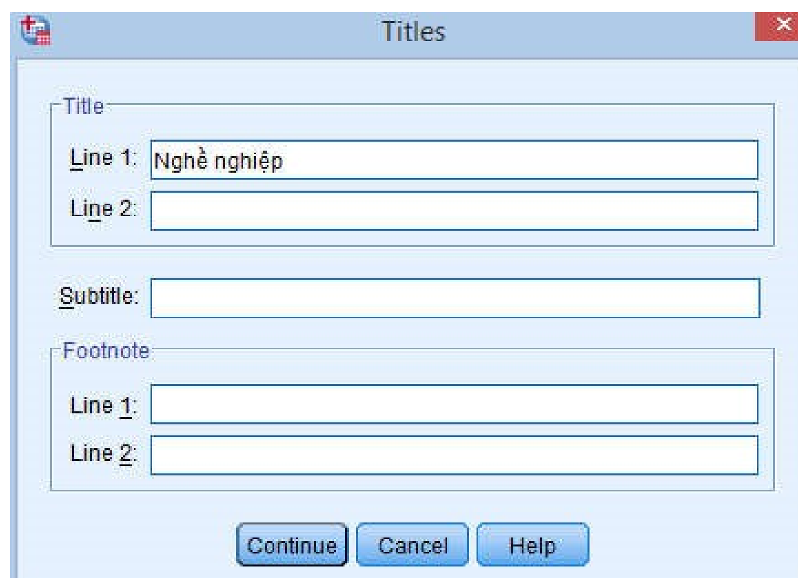
Bước 2. Tại hộp thoại **Bar Charts**, chọn **Clustered**. Tại khu vực **Data in Chart Area** chọn **Summary for group of cases**. Ấn **Define**. Xuất hiện cửa sổ



Hình 26

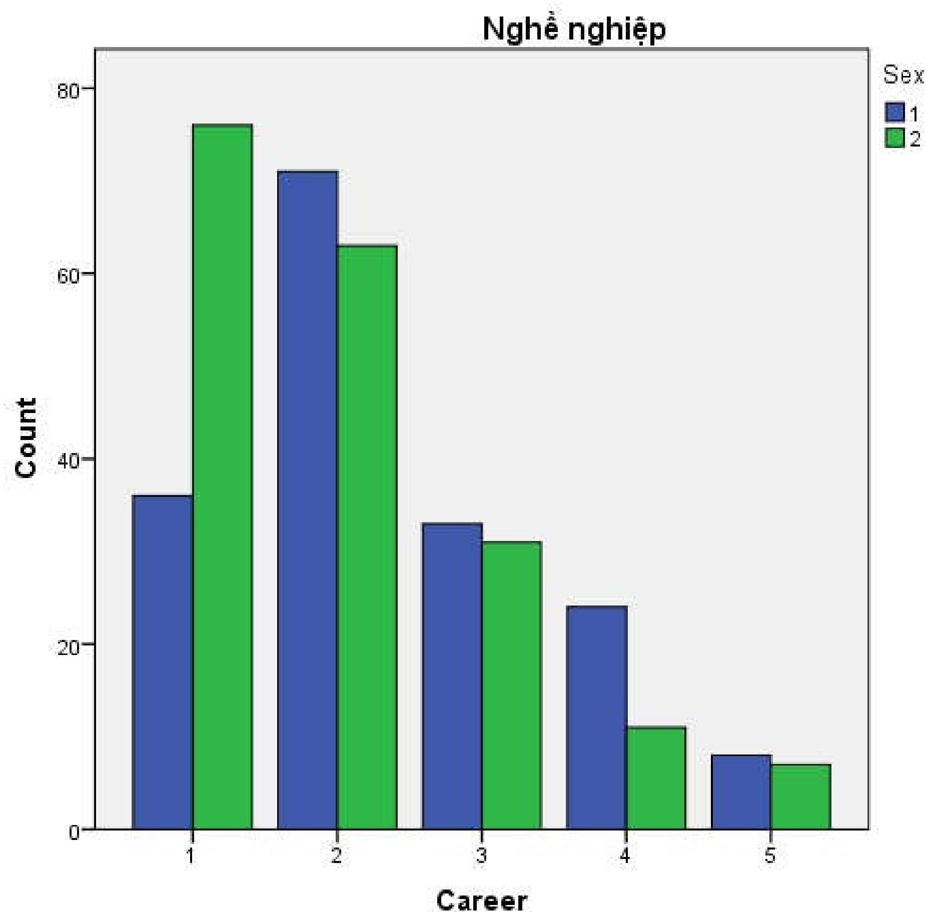
Bước 3. Đưa biến Career vào khung **Category Axis.** Lựa chọn **N of cases** để trục tung biểu diễn tần số, % **of cases** để trục tung biểu diễn tần suất.

Bước 4. Ấn nút **Title** để đặt tiêu đề cho hình rồi ấn **Continue.**



Hình 27

Bước 5. Ấn **OK**. Ta được



Hình 28

4.3. Biểu đồ hình tròn

+) Biểu đồ tròn được dùng cho dữ liệu định danh hay thứ bậc

+) Mỗi hình quạt biểu diễn tỷ lệ đóng góp của từng biểu hiện.

Cách vẽ biểu đồ tròn

Bước 1. Vào menu **Graph Pie** để mở cửa sổ **Pie Chart**

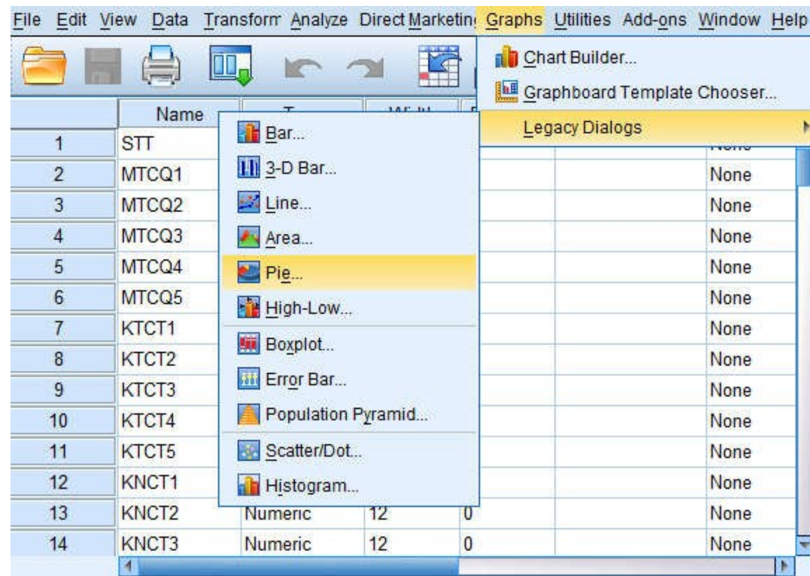
Bước 2. Tại cửa sổ **Pie Chart** chọn **Summary for group of cases**. Ấn **Define**. Xuất hiện cửa sổ sau

Bước 3. Đưa biến cần biểu diễn vào khung **Define Slice by**. Ấn **OK**

Ví dụ: vẽ biểu đồ hình tròn biểu diễn một biến

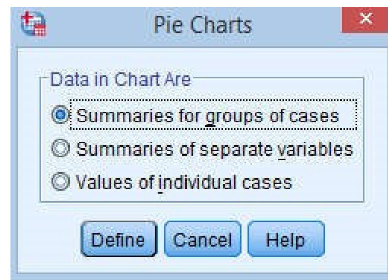
Để vẽ biểu đồ thanh biểu diễn biến nghề nghiệp (Career) trong file **Vidu2.sav** ta tiến hành như sau

Bước 1. Từ cửa sổ **Data View Graphs Legacy Dialogs Pie**



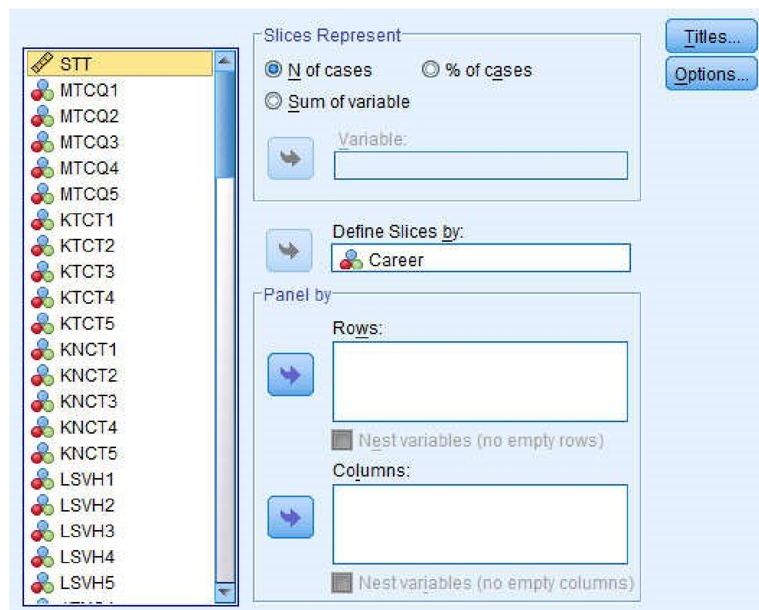
Hình 29

Kích đúp **Pie**, ta được



Hình 30

Bước 2. Tại cửa sổ **Pie Chart** chọn **Summary for group of cases**. Ấn **Define**. Xuất hiện cửa sổ sau



Hình 31

Bước 3. Đưa biến **Career** vào khung **Define Slice by**.

Bước 4. Ấn nút **Title** để đặt tiêu đề cho hình rồi ấn **Continue.**

Titles

Title

Line 1: Nghề nghiệp

Line 2:

Subtitle:

Footnote

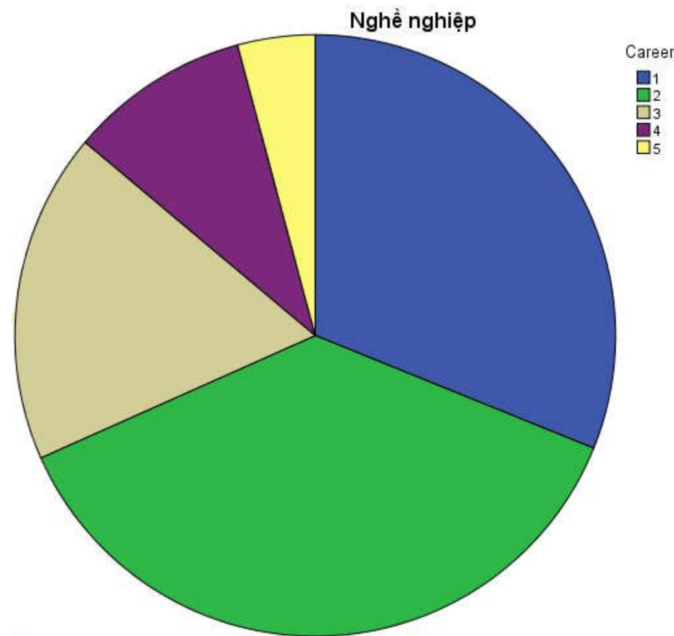
Line 1:

Line 2:

Continue Cancel Help

Hình 32

Bước 5. Ấn **OK**. Ta được



Hình 33

4.4. Biểu đồ tần số (Histogram)

+) Biểu đồ tần số: dùng cho biến định lượng liên tục nhằm biểu diễn phân phối của tập dữ liệu.

+) Chiều cao mỗi thanh hình chữ nhật trong biểu đồ biểu diễn tần số của lớp tương ứng.

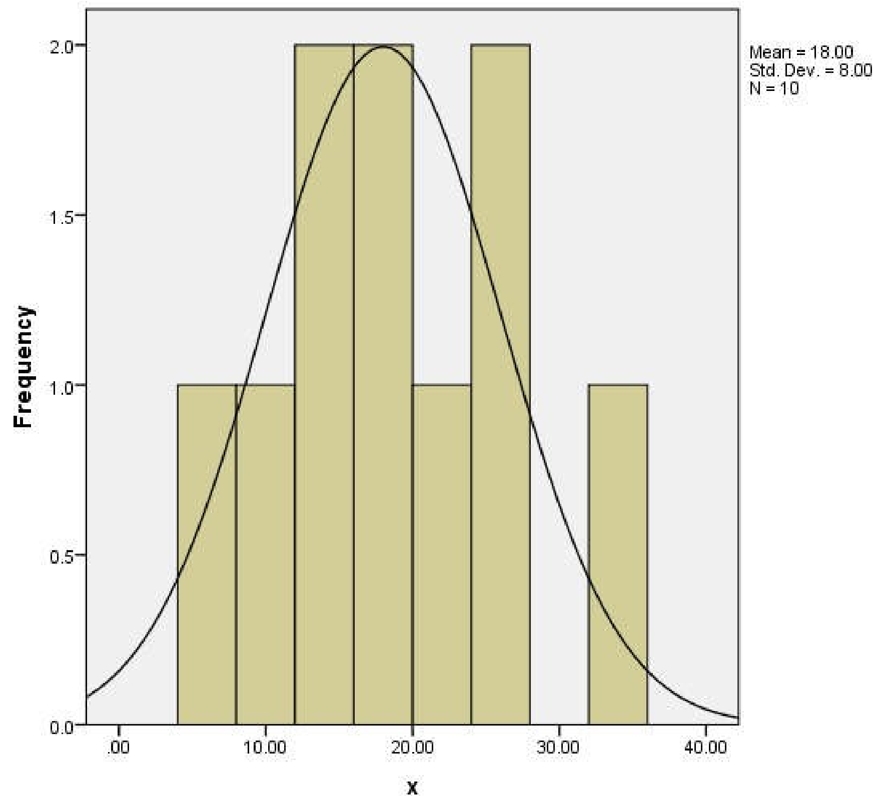
Cách vẽ biểu đồ tần số (Histogram)

Bước 1. Chọn **Graph□Histogram**

Bước 2. Chọn một biến định lượng vào khung **Variable**

Bước 3. Nếu muốn hiển thị đường phân phối chuẩn thì chọn **Display normal curve**. Ấn **OK**.

Ví dụ ta cần vẽ biểu đồ tần số cho biến tuổi trong file **Vidu1.sav**. Trong bước hai ta chọn biến định lượng là X. Hình vẽ như sau



Hình 34

4.5. Biểu đồ phân tán (Scatter plot)

Biểu đồ tán xạ dùng để biểu diễn mối quan hệ giữa hai biến định lượng.

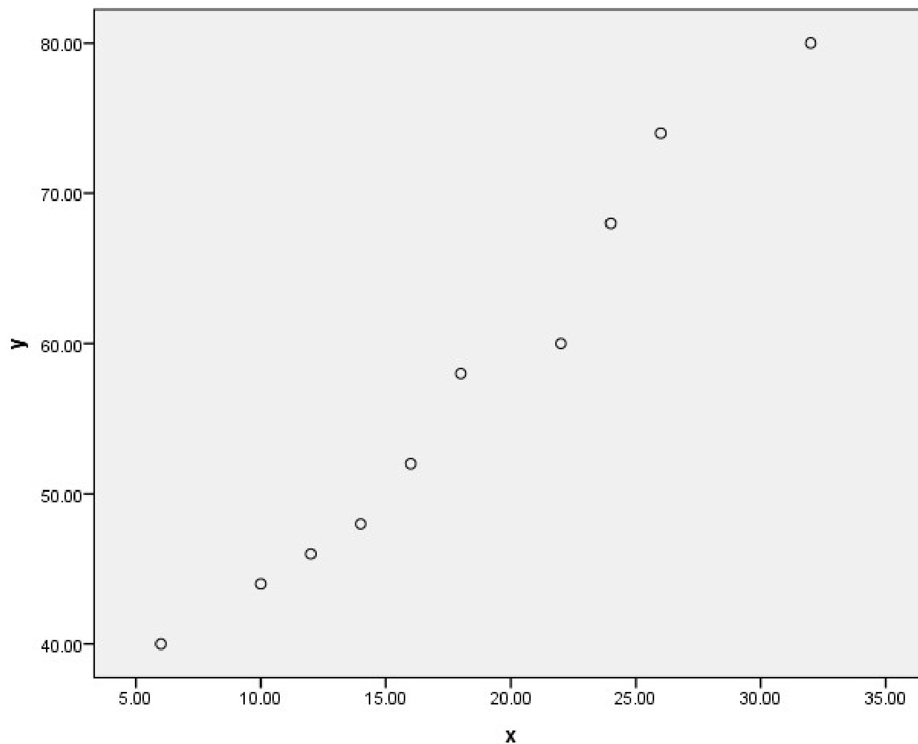
Cách vẽ biểu đồ phân tán (Scatter plot)

Bước 1: Chọn **GraphScatter/Dot**. Được cửa sổ sau

Bước 2: Tại cửa sổ **Scatter/Dot**, chọn **Simple** nếu vẽ biểu đồ cho một cặp biến. (**Ovelay** biểu diễn nhiều đám mây đơn cùng biểu đồ,....)

Bước 3: Chọn biến đưa vào khung **Y-axis** (giá trị là tung độ của điểm), biến đưa vào khung **X-axis** (giá trị là hoành độ của điểm)

Ví dụ ta cần vẽ biểu đồ phân tán trong file **Vidu1.sav**. Hình vẽ như sau

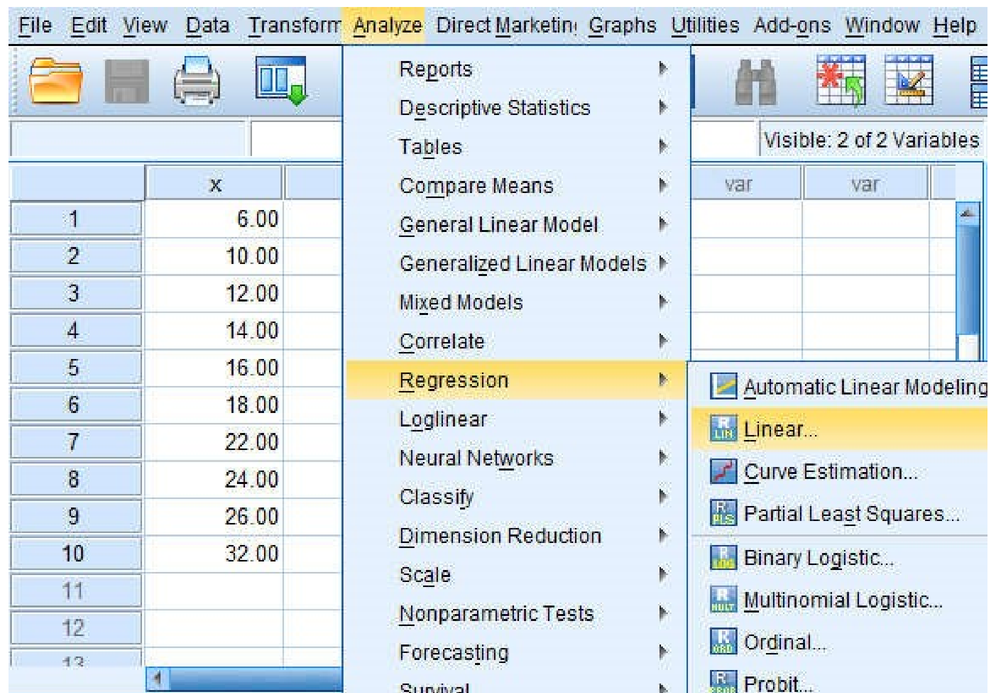


Hình 35

5. Tìm hàm hồi quy tuyến tính, kiểm định mô hình, ước lượng hệ số hồi quy

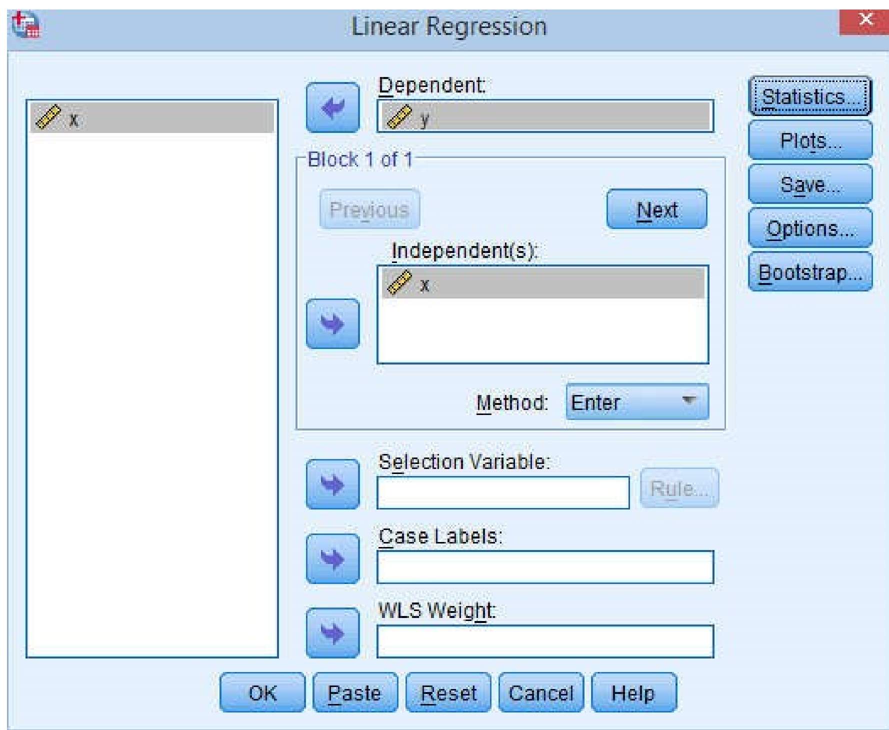
Ta cần tìm hàm hồi quy tuyến tính, kiểm định mô hình, ước lượng hệ số hồi quy trong file **Vidu1.sav**. Ta thực hiện như sau

Bước 1. Từ cửa sổ **Data View** chọn **Analyze Regression Linear**



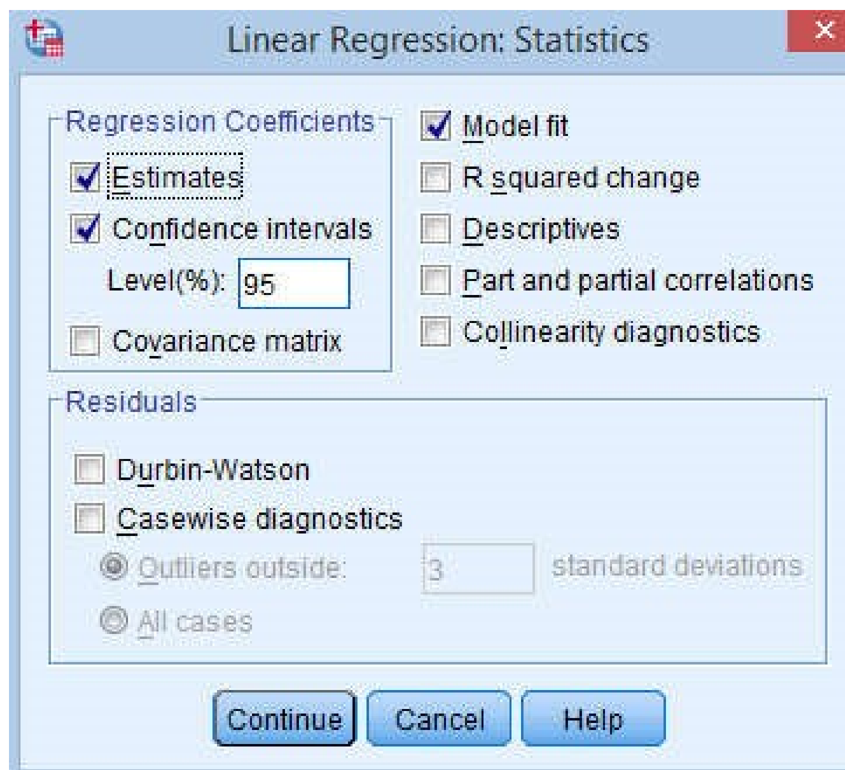
Hình 36

Kích đúp **Linear,** ta được



Hình 37

Bước 2. Đưa biến Y sau **Dependent** và biến X **vào Independent(s).** Chọn **Statistics** ta được bảng sau



Hình 38

Tích vào **Confidence intervals**. Ấn **Continue**

Kích OK, ta được các bảng sau

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.985[a]	.971	.967	2.43171

a. Predictors: (Constant), x

Bảng 6. Tóm tắt mô hình

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1586.694	1	1586.694	268.331	.000[b]
	Residual	47.306	8	5.913		
	Total	1634.000	9			

a. Dependent Variable: y

b. Predictors: (Constant), x

Bảng 7. Phân tích phương sai

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	27.125	1.979		13.705	.000	22.561	31.689
	x	1.660	.101	.985	16.381	.000	1.426	1.893

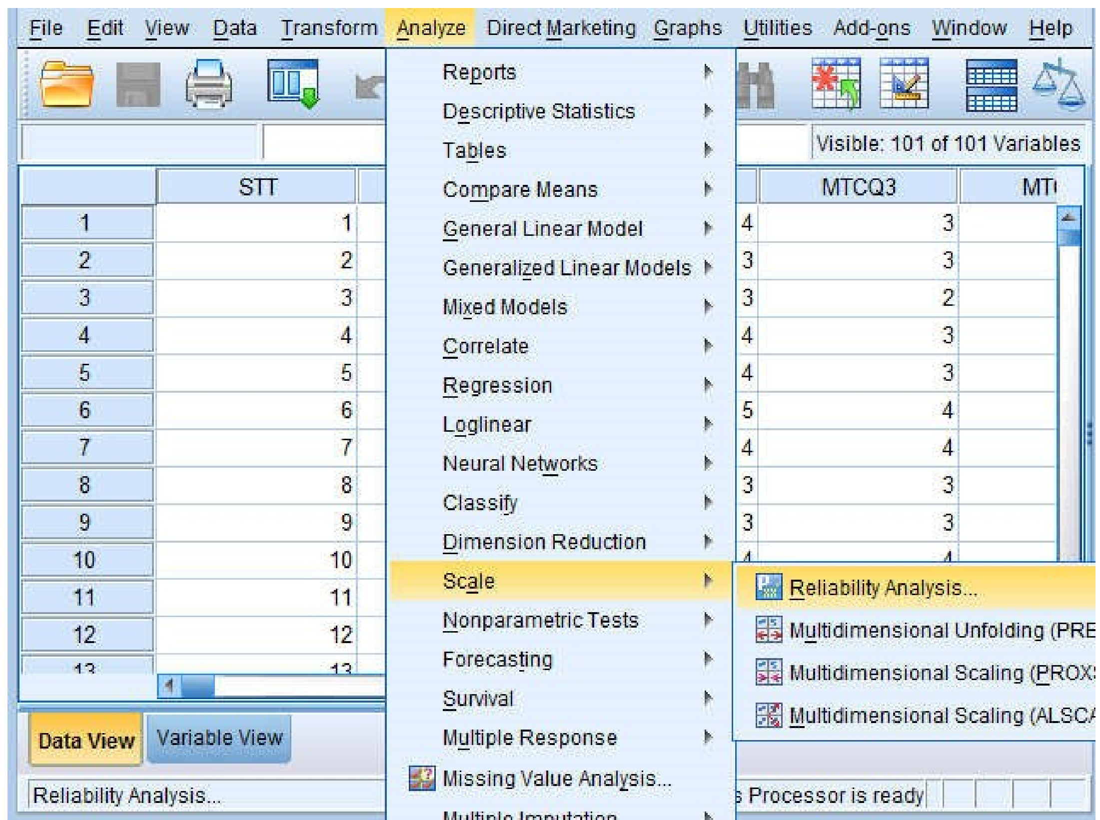
a. Dependent Variable: y

Bảng 8. Ước lượng hệ số hồi quy

6. Đánh giá độ tin cậy các thang đo bằng Cronbach 's alpha

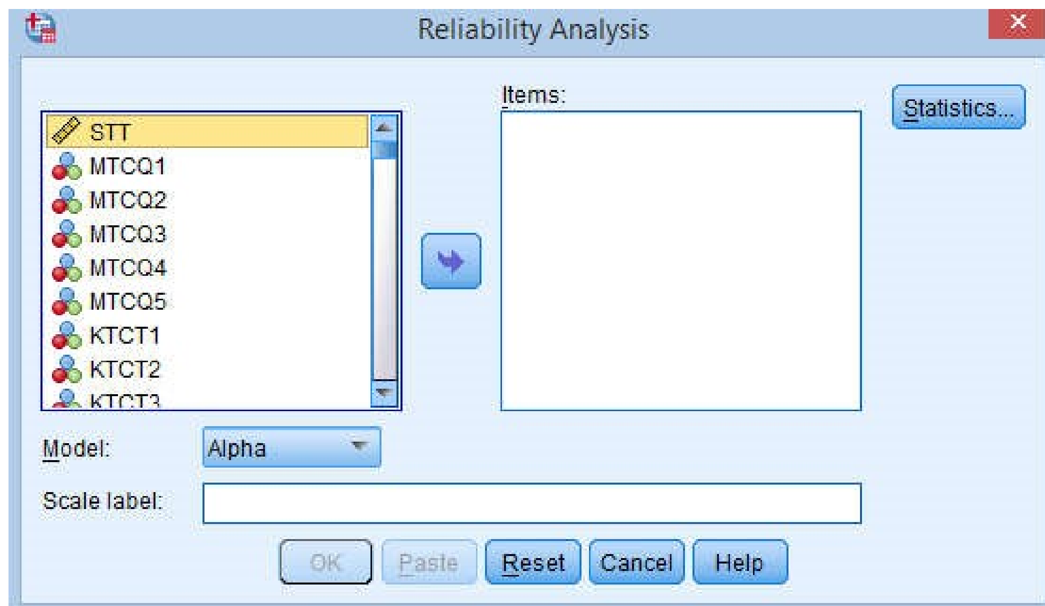
Trong phần này áp dụng cụ thể cho file **Vidu2.sav**. Ta kiểm định thang đo của môi trường cảnh quan (MTCQ1,…, MTCQ5) như sau

Từ cửa số **Data View:** Chọn **Analyze□Scale□Reliability Analysis…**như sau:



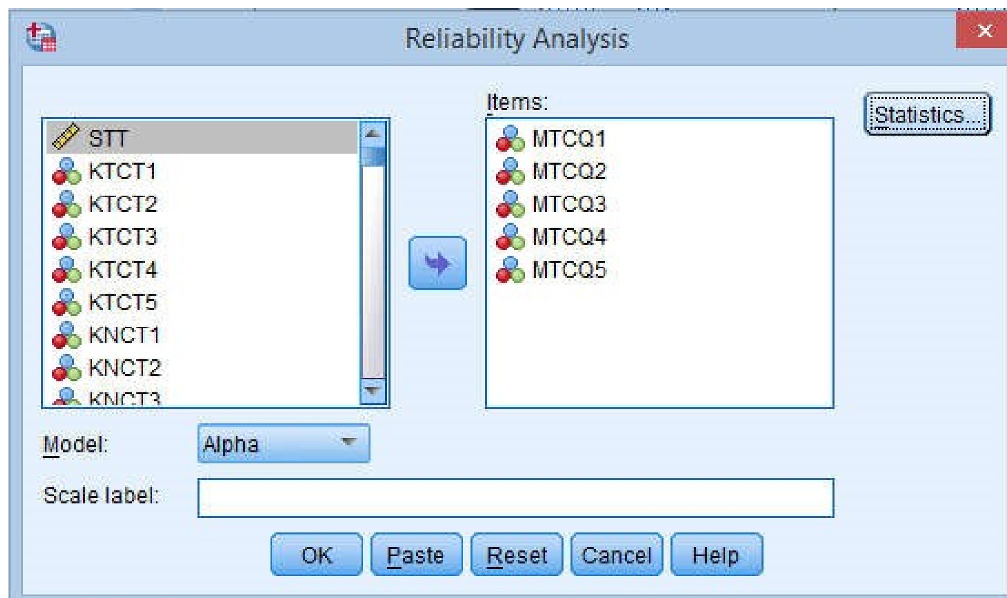
Hình 39

Kích đúp **Reliability Analysis,** ta được



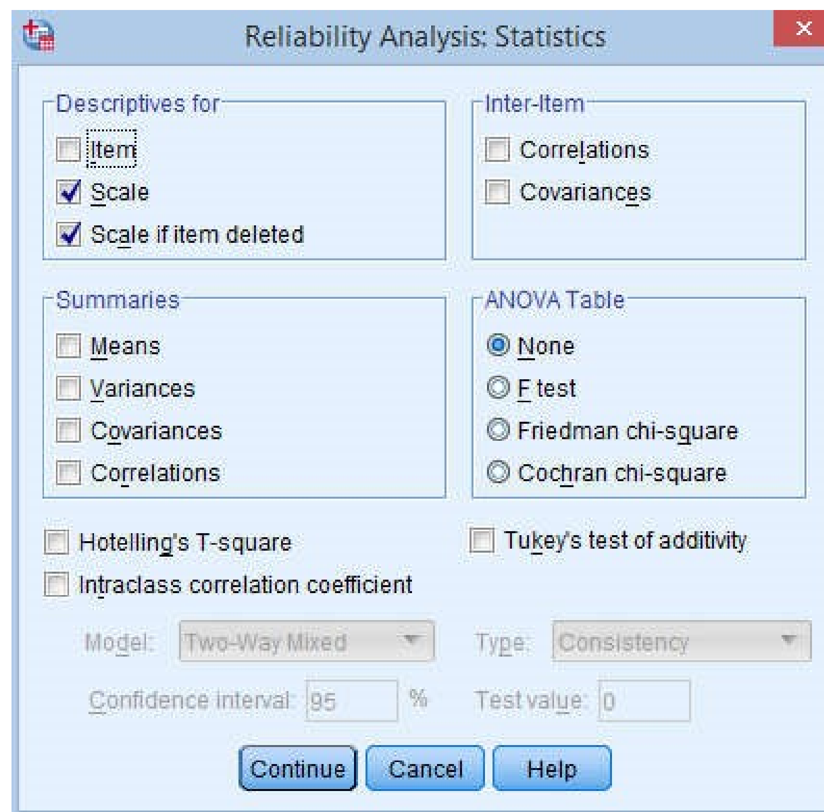
Hình 40

Trong hình 40, chọn nhóm biến môi trường cảnh quan (**MTCQ1,…, MTCQ5**) đưa vào hộp **Item**, xuất hiện như sau:



Hình 41

Trong hình 41. Chọn **Statistics**, màn hình xuất hiện như sau:



Hình 42

Trong hình 42, chọn **Scale** và **Scale if Scale item deleted** của hộp **Descriptives for**, sau đó chọn **Continue** màn hình xuất hiện như hình 41. Chọn **OK**, ta được

Reliability Statistics

Cronbach's Alpha	N of Items
.742	5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
MTCQ1	14.11	6.195	.421	.730
MTCQ2	13.96	5.823	.598	.663
MTCQ3	14.02	5.668	.553	.678
MTCQ4	13.99	6.237	.485	.705
MTCQ5	13.99	6.248	.479	.707

Bảng 9

Sau khi chạy Cronbach ‘s alpha từng yếu tố, biến KNTC1 bị loại vì hệ số tương quan biến - tổng (Corrected Item – total Correlation) là -0,138 (<0,3).

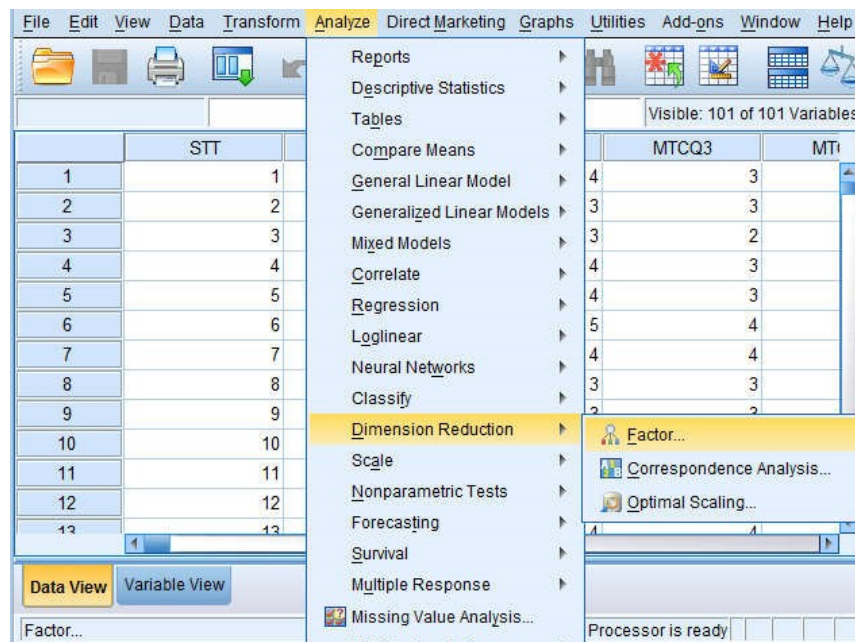
Kết quả đánh giá Cronbach ‘s alpha cho thấy các thang đo đều có hệ số alpha đạt yêu cầu (> 0,6), thấp nhất là thang đo khả năng tiếp cận điểm đến (KNTC) có alpha = 0,638 và cao nhất là thang đo cơ sở hạ tầng (CSHT) có alpha = 0,836.

7. Phân tích nhân tố khám phá (EFA)

Trong phần này áp dụng cụ thể cho file **Vidu2.sav**. Ta tiến hành như sau

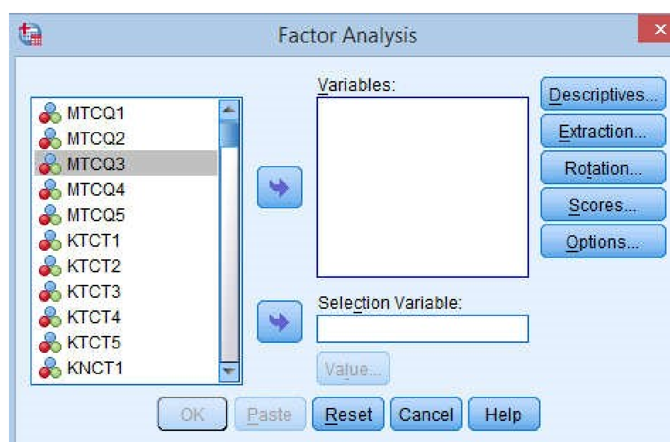
7.1. Sử dụng SPSS thực hiện phân tích EFA

Từ cửa sổ **Data View:** Chọn **Analyze□Dimension Reduction□Factor…**như sau



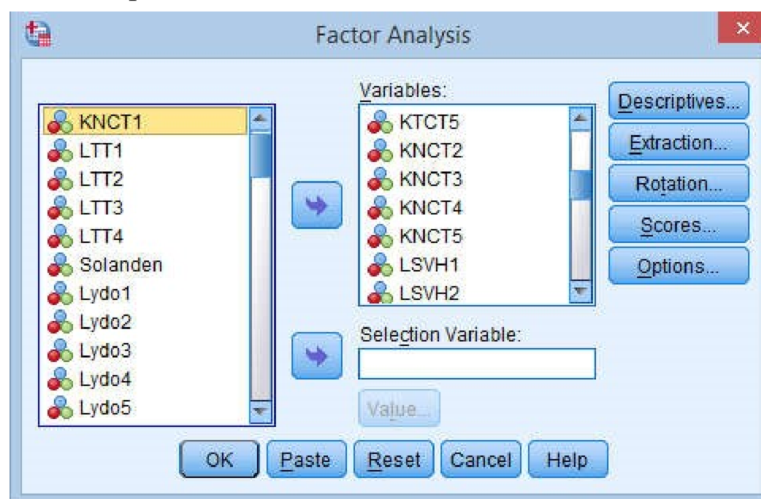
Hình 43

Kích đúp **Factor,** ta được



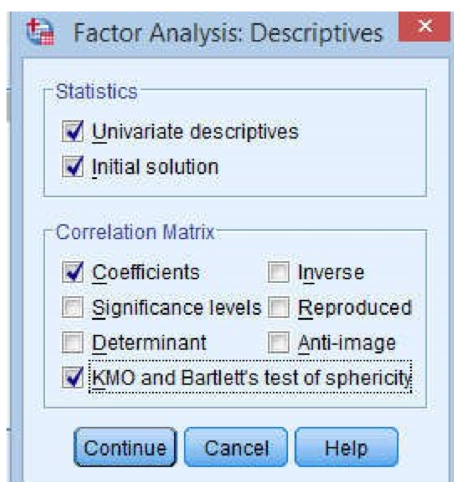
Hình 44

Trong hình 44, đưa tất cả các biến đặc trưng vào hộp **Variables** (trừ biến bị loại ở kiểm định Cronbach's Alpha và các biến của thước đo **LLT**, xuất hiện như sau:



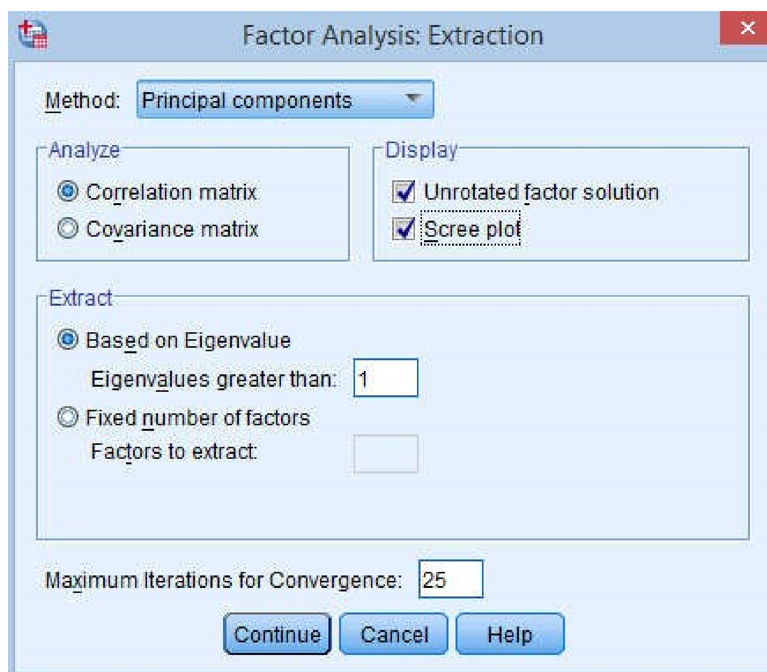
Hình 45

Chọn **Descriptives**, sẽ xuất hiện



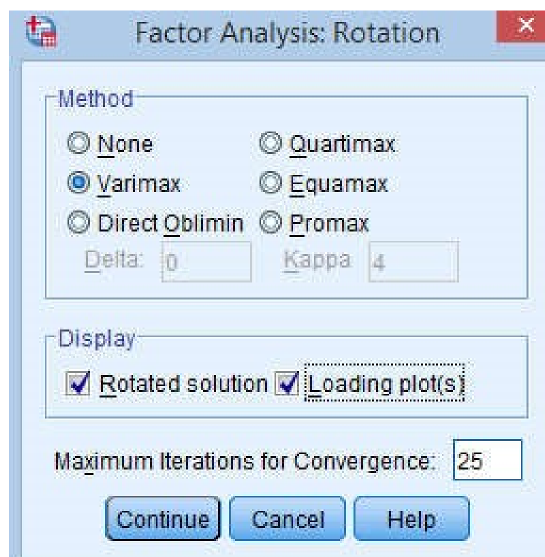
Hình 46

Trong hình 46, ta chọn **Univariate descriptives** và **Initial solution** trong hộp **Statistics. Trong hộp Correlation Matrix, chọn Coefficients và KMO.** Nhấn **Continue**, chọn **Extraction** sẽ xuất hiện



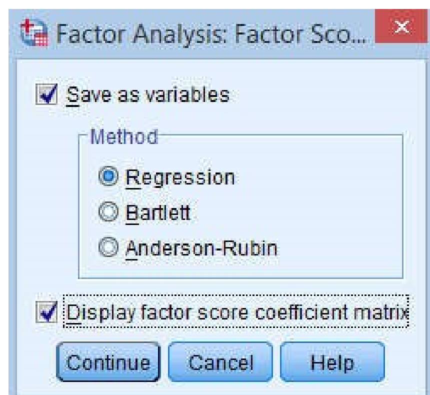
Hình 47

Trong hình 47, ta chọn **Principle components** trong hộp **Method**. Trong hộp **Analyze**, chọn **correlation matrix**. Trong hộp **Display,** chọn **Unrotated factor solution** và **Scree plot**. Trong hộp **Extract**, chọn **Based on Eigenvalue** và **Eigenvalues greater than 1**. Trong hộp **Maximum Iterations for Convergence 25**. Nhấn **Continue**, chọn **Rotation** sẽ xuất hiện



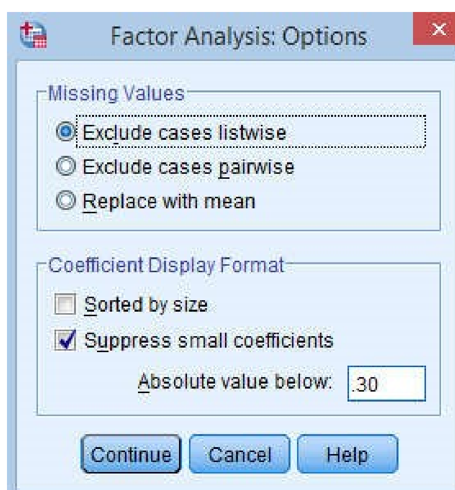
Hình 48

Trong hình 48, ta chọn **Varimax** trong hộp **Method**. Trong hộp **Display**, chọn **Rotated solution** và **Losding plot(s).** Trong hộp **Maximum Iterations for Convergence 25**. Nhấn **Continue**, chọn **Scores** sẽ xuất hiện



Hình 49

Trong hình 49, bên dưới **Factor Analysis**, ta chọn **Save as variables** và **Display factor score coefficient matrix**. Trong hộp **Method,** chọn **Regression**. Nhấn **Continue**, chọn **Option** sẽ xuất hiện



Hình 50

Trong hình 50, chọn **Exclude cases listwise**. Trong hộp **Coefficient Display Format,** chọn **Suppress small coefficient.** Đối với **Absolute value Below,** chọn giá trị bao nhiêu tùy thuộc vào cỡ mẫu nghiên cứu. **Lưu ý**

Cỡ mẫu	**Giá trị chọn (Factor loadings)**
Lớn hơn 350	0,3
Từ 100 đến 350	0,55
Nhỏ hơn 100	0,75

Bảng 10

Trong ví dụ trên cỡ mẫu 360 ta chọn 0,3.

Nhấn **Continue**, chọn **OK** ta có toàn bộ kết quả

Factor Analysis

[DataSet1] C:\Users\User\Desktop\luanvancongoc.sav

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
MTCQ1	3.41	.891	360
MTCQ2	3.56	.819	360
MTCQ3	3.49	.902	360
MTCQ4	3.53	.810	360
MTCQ5	3.53	.814	360
KTCT1	3.72	.783	360
KTCT2	3.98	.763	360
KTCT3	3.52	.824	360
KTCT4	3.72	.733	360
KTCT5	3.82	.722	360
KNCT2	3.75	1.058	360
KNCT3	3.79	.939	360
KNCT4	3.66	1.043	360

Bảng 11

Rotated Component Matrix[a]

	Component						
	1	2	3	4	5	6	7
MTCQ1					.606		
MTCQ2					.777		
MTCQ3					.754		
MTCQ4					.691		
MTCQ5					.667		
KTCT1		.699					
KTCT2		.763					
KTCT3		.770					
KTCT4		.754					
KTCT5		.678					
KNCT2							.758
KNCT3							.713
KNCT4							.743
LSVH1			.640				
LSVH2			.711				
LSVH3			.768				
LSVH4			.682				
LSVH5			.704				

Bảng 12. Kết quả phân tích nhân tố khám phá

Ta chọn các kết quả cần thiết phục vụ cho phân tích.

7.2. Hệ thống kiểm định cho EFA

7.2.1. Kiểm định tính thích hợp của EFA

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy. Approx. Chi-Square Bartlett's Test of Sphericity df Sig.	.759 3239.022 496 .000

Bảng 13 . Kiểm định KMO và Bartlett

Dựa vào kết quả ta có hệ số KMO là 0,759 (>0,5) và sig= 0,00 <0,05, nghĩa là các biến quan sát có tương quan tuyến tính với nhân tố đại diện.

7.2.2. Kiểm định mức độ giải thích của các biến quan sát với nhân tố

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulativ e %
1	3.868	12.087	12.087	3.868	12.087	12.087	3.090	9.655	9.655
2	3.337	10.429	22.515	3.337	10.429	22.515	2.906	9.082	18.738
3	2.725	8.516	31.032	2.725	8.516	31.032	2.672	8.350	27.087
4	2.316	7.237	38.268	2.316	7.237	38.268	2.564	8.013	35.100
5	2.151	6.723	44.991	2.151	6.723	44.991	2.512	7.851	42.951
6	1.821	5.690	50.681	1.821	5.690	50.681	2.252	7.038	49.989
7	1.524	4.762	55.443	1.524	4.762	55.443	1.745	5.454	55.443
8	.973	3.039	58.482						
9	.961	3.002	61.484						
10	.808	2.526	64.010						

Bảng 14. Tổng phương sai được giải thích.

Phân tích EFA mang lại 7 nhóm nhân tố tác động đến LLT. Hệ số tải và hệ số tin cậy Cronbach Alpha đạt tại Eigenvalue = 1,524 và phương sai trích đạt 55,443 % (> 0,50). Vì thế kết quả EFA là đáng tin cậy và có thể sử dụng cho phân tích hồi qui ở bước theo.

7.2.3. Kết quả mô hình EFA Rotated

Component Matrix[a]

	Component						
	1	2	3	4	5	6	7
MTCQ1					.606		
MTCQ2					.777		

MTCQ3					.754		
MTCQ4					.691		
MTCQ5					.667		
KTCT1		.699					
KTCT2		.763					
KTCT3		.770					
KTCT4		.754					
KTCT5		.678					
KNCT2							.758
KNCT3							.713
KNCT4							.743
LSVH1			.640				
LSVH2			.711				
LSVH3			.768				
LSVH4			.682				
LSVH5			.704				
ATMS1				.732			
ATMS2				.758			
ATMS3				.718			
ATMS4				.670			
ATMS5				.622			
CSHT1	.762						
CSHT2	.826						
CSHT3	.809						
CSHT4	.749						
CSHT5	.717						
GT1						.788	
GT2						.728	
GT3						.760	
GT4						.695	

Bảng 15. Ma trận nhân tố xoay.

Trong bảng 15. Cho biết các đặc trưng đều có hệ số tải nhân tố lớn hơn 0,5. Có 7 nhân tố

+) Nhân tố 1. Bao gồm các biến CSHT2, CSHT3, CSHT1, CSHT4, CSHT5. Ta đặt tên cho nhân tố này là **CSHT (F1).**

+) Nhân tố 2. Bao gồm các biến KTCT3, KTCT2, KTCT4, KTCT1, KTCT5. Ta đặt tên cho nhân tố này là **KTCT (F2).**

+) Nhân tố 3. Bao gồm các biến LSVH3, LSVH2, LSVH5, LSVH4, LSVH1. Ta đặt tên cho nhân tố này là **LSVH (F3).**

+) Nhân tố 4. Bao gồm các biến ATMS2, ATMS1, ATMS3, ATMS4, ATMS5. Ta đặt tên cho nhân tố này là **ATMS (F4).**

+) Nhân tố 5. Bao gồm các biến MTCQ2, MTCQ3, MTCQ4, MTCQ5, MTCQ1. Ta đặt tên cho nhân tố này là **MTCQ (F5).**

+) Nhân tố 6. Bao gồm các biến GT1, GT3, GT2, GT4. Ta đặt tên cho nhân tố này là **GT (F6).**

+) Nhân tố 7. Bao gồm các biến KNCT2, KNCT4, KNCT3. Ta đặt tên cho nhân tố này là **KNCT (F7).**

7.2.4. Thực hiện phân tích hồi quy đa biến

Để nhận diện các yếu tố ảnh hưởng đến hình ảnh điểm đến, mô hình tương quan tổng thể có dạng:

$$LTT = f(F_1, F_2, F_3, F_4, F_5, F_6, F_7)$$

Trong đó

+) LTT là biến phụ thuộc;

+) F1, F2, F3, F4, F5, F6, F7 là biến độc lập.

Vì thế, chúng ta có thể dự đoán mô hình hồi qui bội với 7 nhân tố như đã xác định trong các chương trước có dạng như sau:

$$LTT = \beta_0 + \beta_1 F_1 + \beta_2 F_2 + \beta_3 F_3 + \beta_4 F_4 + \beta_5 F_5 + \beta_6 F_6 + \beta_7 F_7$$

Trong đó các biến đưa vào phân tích hồi quy được xác định bằng cách tính điểm của các nhân tố (Factor score).

Nhân tố thứ i, được xác định

$$F_i = w_{i1} X_1 + w_{i2} X_2 + \ldots + w_{ik} X_k$$

Với

+) w_{ik} : Hệ số nhân tố được trình bày trong ma trận hệ số nhân tố (Component Score Coefficient).

+) X_i: Biến quan sát trong nhận tố i.

Đối với các biến độc lập, SPSS tính sẵn trong Data view khi phân tích nhân tố khám phá như hình 51

*Untitled4 [DataSet3] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

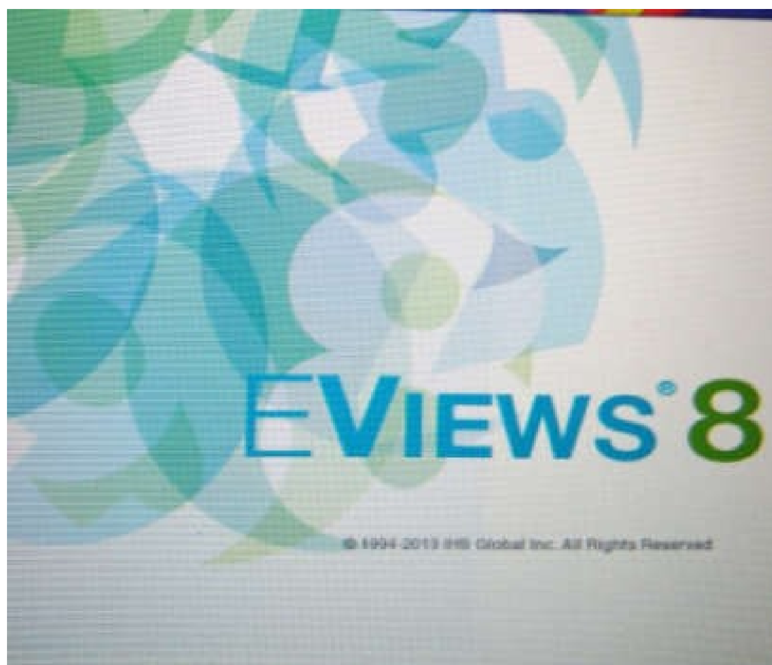
Visible: 107 of 107 Variables

	FAC1_1	FAC2_1	FAC3_1	FAC4_1
1	.40822	.31880	1.01112	.57967
2	.10123	1.05695	-.22270	-1.28885
3	-1.63966	1.06304	.56903	-1.40693
4	-1.55669	.23646	.76974	1.05450
5	.51923	1.12725	.11153	.85593
6	-.68025	1.42337	-.55591	-.02590
7	-.75984	-.86906	-.53831	.61134
8	-2.02780	.59018	-.15160	-1.20010
9	-.00507	.30466	.03281	-.11535
10	.28203	.56181	.80585	1.32571
11	.28010	1.37532	2.05093	.67201
12	-.11275	.25307	-.75395	.08857
13	.40206	.47069	.24006	.98813

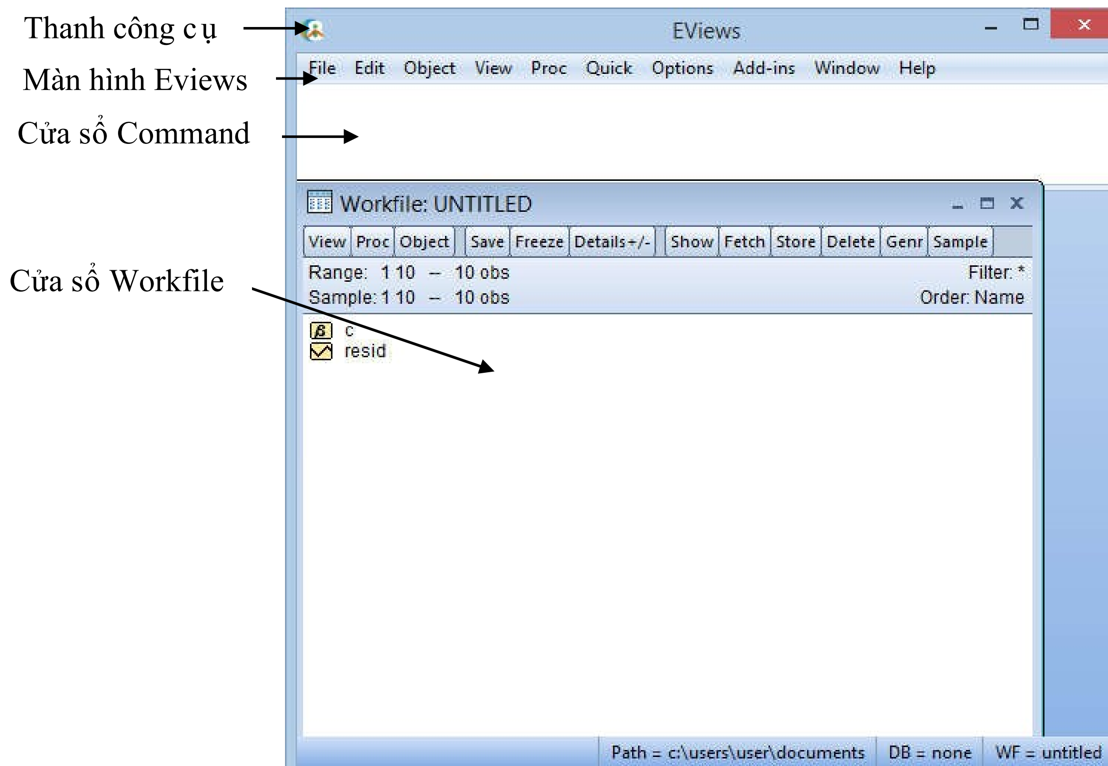
Hình 51

Trong hình 51, FAC1_1 là F1, FAC2_1 là F3, FAC3_1 là F3, FAC4_1 là F4, được xác định theo cách tính điểm của nhân tố, tương tự cho các nhân tố còn lại.

Phụ lục 3. HƯỚNG DẪN SỬ DỤNG PHẦN MỀM EVIEW 8.0



1. Màn hình Eviews



Hình 1

2. Các kiểu dữ liệu thường dùng

2.1. Số liệu theo thời gian: là các số liệu thu thập tại nhiều thời điểm khác nhau trên cùng một đối tượng. Chẳng hạn như số liệu về GDP bình quân của Việt Nam từ 1998 – 2006 được cho trong bảng sau:

Năm	1998	1999	2000	2001	2002	2003	2004	2005	2006
GDP	360	374	401	413	440	489	553	618	655

Bảng 1

2.2. Số liệu chéo: là số liệu thu thập tại một thời điểm ở nhiều nơi, địa phương, đơn vị, khác nhau. Chẳng hạn như số liệu về GDP bình quân trong năm 2006 của các nước Brunei, Campuchia, Indonesia, Lào, Malaysia, Myanmar, Philippines, Singapore, Thái Lan, Việt Nam được cho như sau:

Nước	GDP
Brunei	30376
Campuchia	459
Indonesia	1581
Lào	570
Malaysia	5570

Nước	GDP
Myanmar	230
Philippines	1361
Singapore	30162
Thái Lan	2959
Việt Nam	655

Bảng 2

2.3. Số liệu hỗn hợp: là số liệu tổng hợp của hai loại trên, nghĩa là các số liệu thu thập tại nhiều thời điểm khác nhau ở nhiều địa phương, đơn vị khác nhau. Chẳng hạn như số liệu về GDP bình quân của các nước từ 1998 – 2006.

Nước	1998	1999	2000	2001	2002	2003	2004	2005	2006
Brunei	13065	14511	18465	16820	17135	18788	21989	25759	30376
Campuchia	255	281	285	302	317	333	373	430	459
Indonesia	516	746	807	773	928	1100	1176	1283	1581
Lào	255	286	329	326	329	378	432	485	570
Malaysia	3254	3485	3844	3665	3884	4161	4652	5042	5570
Myanmar	134	173	178	129	130	197	199	219	230
Philippines	910	1019	994	914	966	982	1049	1168	1361
Singapore	21009	20909	23075	20724	21210	22157	25345	26839	30162
Thái Lan	1829	1985	1967	1836	1999	2233	2484	2659	2959
Việt Nam	360	374	401	413	440	489	553	618	655

Bảng 3

3. Nhập dữ liệu

3.1. Nhập trực tiếp vào Eview

Để minh họa cho phần này, ta xét các ví dụ sau:

Ví dụ 1. Bảng 4 dưới đây cho biết số liệu về GDP bình quân đầu người của Việt Nam trong các năm 1998 – 2006.

Năm	1998	1999	2000	2001	2002	2003	2004	2005	2006
GDP	360	374	401	413	440	489	553	618	655

Bảng 4

Ví dụ 2. Bảng 5 dưới đây cho biết số liệu về doanh số của một công ty.

Năm	Quý			
	I	II	III	IV
2001	5280	4138	3959	7810
2002	6250	4565	4770	8712
2003	5883	5286	6142	9280
2004	7523	5758	5714	10144
2005	6783	6268	7618	11567

Bảng 5

Ví dụ 3. Bảng 6 dưới đây cho biết số liệu về năng suất (Y, đơn vị tạ/ha) và mức phân bón (X, đơn vị tạ/ha) cho một loại cây trồng tính trên một ha trong 10 năm từ 1988 đến 1997.

Năm	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
X	6	10	12	14	16	18	22	24	26	32
Y	40	44	46	48	52	58	60	68	74	80

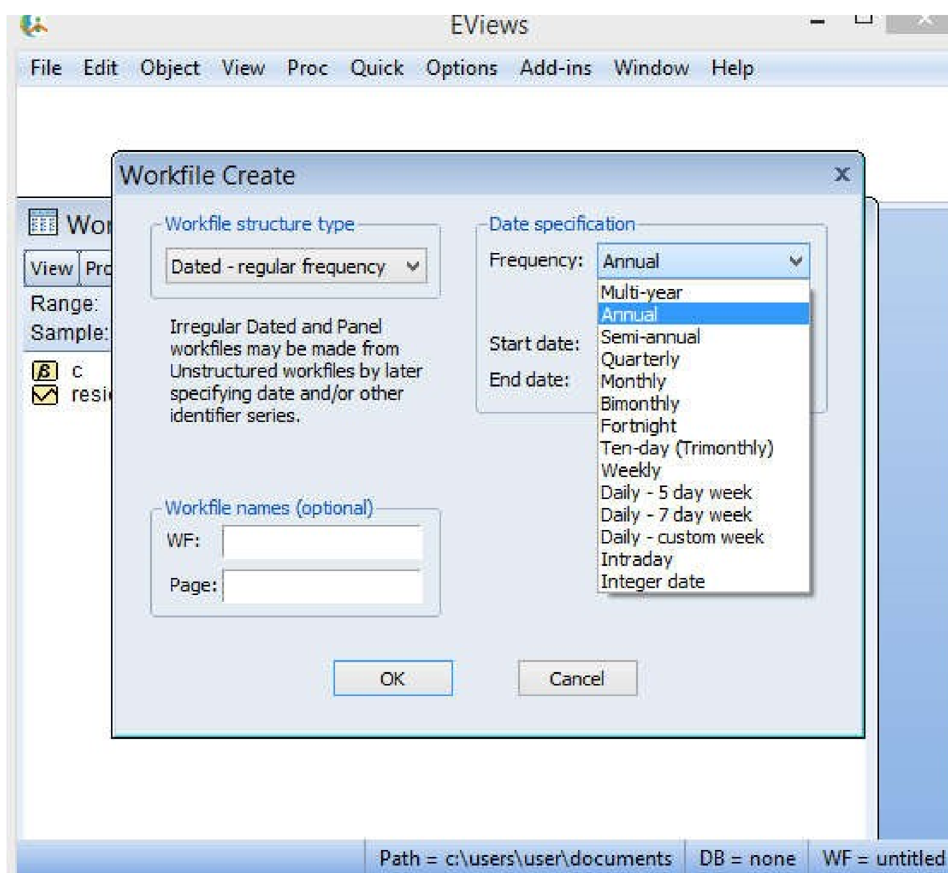
Bảng 6

Ví dụ 4. Bảng 7 dưới đây cho biết số liệu về doanh thu (Y), chi phí cho quảng cáo (X2), tiền lương của nhân viên tiếp thị (X3) của 12 công nhân (đơn vị triệu đồng).

STT	1	2	3	4	5	6	7	8	9	10	11	12
Y_i	127	149	106	163	102	180	161	128	139	144	159	138
X_{2i}	18	25	19	24	15	26	25	16	17	23	22	15
X_{3i}	10	11	6	16	7	17	14	12	12	12	14	15

Bảng 7

Mở Eview, để nhập dữ liệu: Chọn **File□New□Workfile,** ta có màn hình như sau:



Hình 2

Tuỳ vào kiểu dữ liệu cần khảo sát, ta có thể chọn được các kiểu sau :

Dated – regular frequency

Multi – year : Số liệu nhiều năm

Annual : Số liệu năm

Semi – Annual : Số liệu nửa năm

Quarterly : Số liệu theo quý

Monthly : Số liệu theo từng tháng

Bimonthly : Mỗi tháng 2 lần/2 tháng 1 lần

Fortnight : Hai tuần lễ/15 ngày Ten

– day (Trimonthly) :

Weekly : Số liệu theo từng tuần

Unstructure / Undate : Số liệu chéo

Để nhập dữ liệu ở ví dụ 1, ta chọn các khai báo như trong hình 3 như sau:

Workfile Create

Workfile structure type

Dated - regular frequency

Irregular Dated and Panel workfiles may be made from Unstructured workfiles by later specifying date and/or other identifier series.

Date specification

Frequency: Annual

Start date: 1998

End date: 2006

Workfile names (optional)

WF:

Page:

OK Cancel

Hình 3

Để nhập dữ liệu ở ví dụ 2, ta chọn các khai báo như trong hình 4

Workfile Create

Workfile structure type

Dated - regular frequency

Irregular Dated and Panel workfiles may be made from Unstructured workfiles by later specifying date and/or other identifier series.

Date specification

Frequency: Quarterly

Start date: 2001:1

End date: 2005:4

Workfile names (optional)

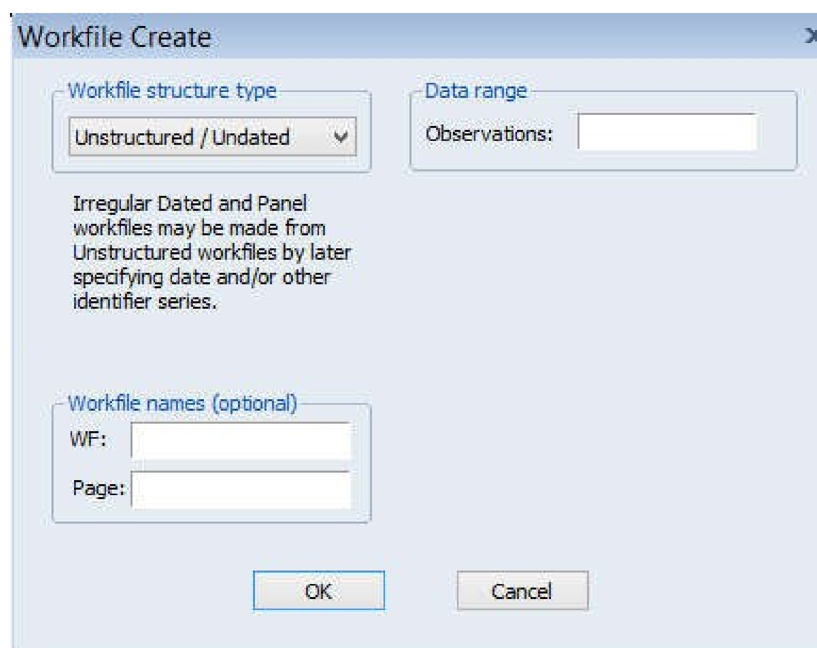
WF:

Page:

OK Cancel

Hình 4

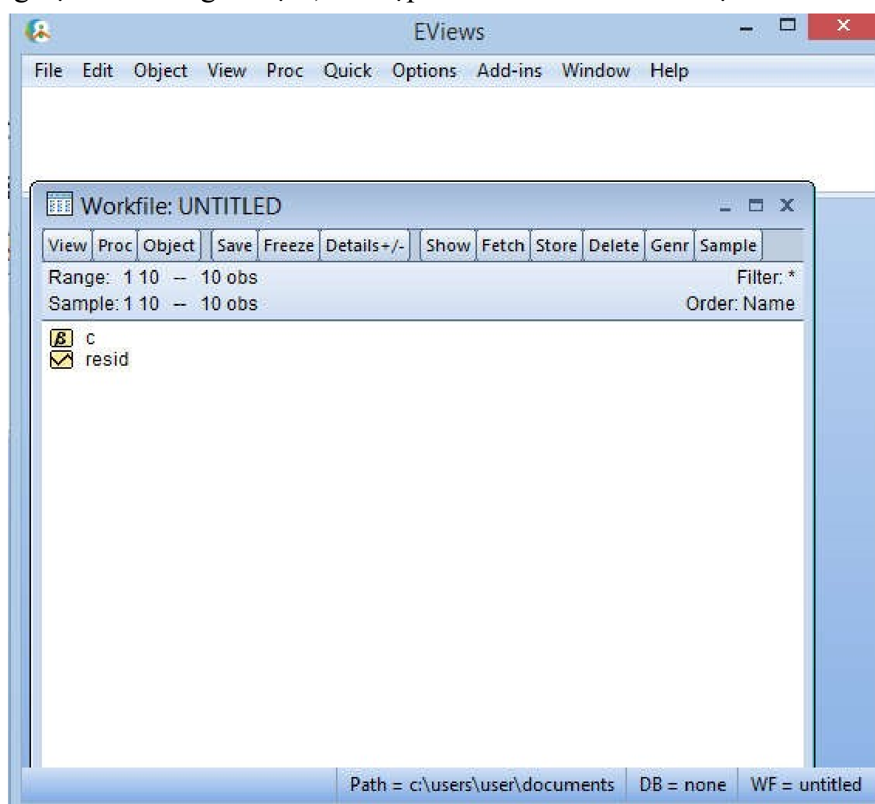
Để nhập dữ liệu cho ví dụ 3, ví dụ 4, ta có thể khai báo báo như trong hình 5.



Hình 5

Trong ô **Observations** ta nhập cỡ mẫu (số các quan sát)

Chẳng hạn như trong ví dụ 3, ta nhập 10 rồi **nhấn OK** ta được hình 6



Hình 6

Để nhập số liệu ta chọn : **Quick →Empty Group (Edit Series)**, màn hình xuất hiện một cửa sổ như hình 7. Trong đó

- Cột **obs** ghi thứ tự quan sát.
- Các cột kế tiếp để khai báo các biến và nhập số liệu.

G Group: UNTITLED Workfile: UNTITLED::Untitled\

View | Proc | Object | Print | Name | Freeze | Default | Sort | Edit+/- | Smpl+/- | Compare+/-

1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

Hình 7

Ví dụ nhập số liệu cho biến Y vào cột số 2, ta nhấp chuột vào đầu cột này và gõ tên biến Y sau đó nhấp Enter và lần lượt gõ các giá trị vào các ô bên dưới có ghi chữ NA. Chẳng hạn như trong ví du 3 và ví dụ 4, ta khai báo và nhập số liệu tuần tự như trong các hình sau :

G Group: UNTITLED Workfile: UNTITLED::Untitled\

View | Proc | Object | Print | Name | Freeze | Default | Sort | Edit+/- | Smpl+/- | Compare+/-

	X	Y		
	X	Y		
1	6.000000	40.00000		
2	10.00000	44.00000		
3	12.00000	46.00000		
4	14.00000	48.00000		
5	16.00000	52.00000		
6	18.00000	58.00000		
7	22.00000	60.00000		
8	24.00000	68.00000		
9	26.00000	74.00000		
10	32.00000	80.00000		

Hình 8

Group: UNTITLED Workfile: UNTITLED::Untitled\

View | Proc | Object | Print | Name | Freeze | Default | Sort | Edit+/- | Smpl+/- | Compare+/-

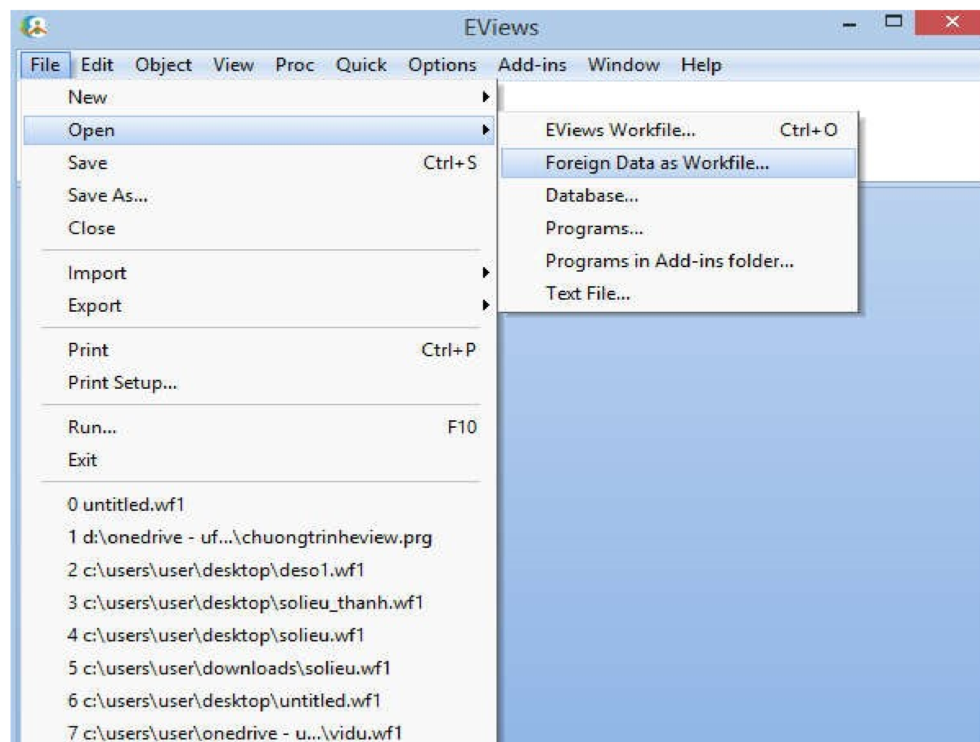
	Y	X2	X3
	Y	X2	X3
1	127.0000	18.00000	10.00000
2	149.0000	25.00000	11.00000
3	106.0000	19.00000	6.000000
4	163.0000	24.00000	16.00000
5	102.0000	15.00000	7.000000
6	180.0000	26.00000	17.00000
7	161.0000	25.00000	14.00000
8	128.0000	16.00000	12.00000
9	139.0000	17.00000	12.00000
10	144.0000	23.00000	12.00000
11	159.0000	22.00000	14.00000
12	138.0000	15.00000	15.00000

Hình 9

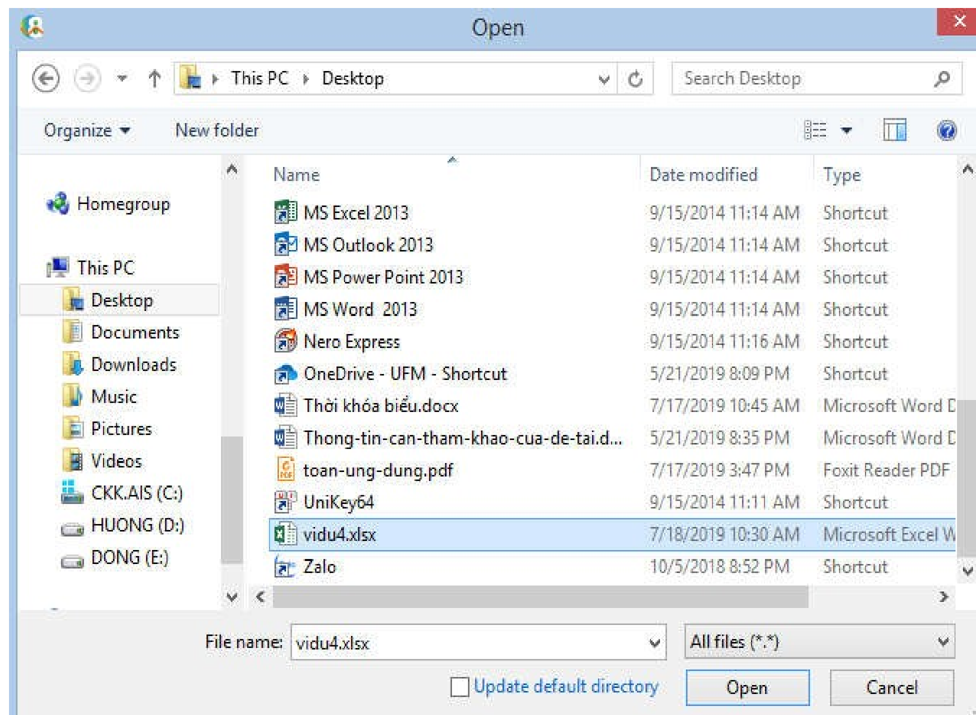
3.2. Nhập từ Excel và Word có sẵn

Giả sử ta có sẵn File Excel **vidu 4.xls** chứa số liệu của ví dụ 4. Khi đó ta thực hiện các bước Import sau:

Mở chương trình **Eviews** chọn **File** → **Open** →**Foreign Data as Workfile**…như sau

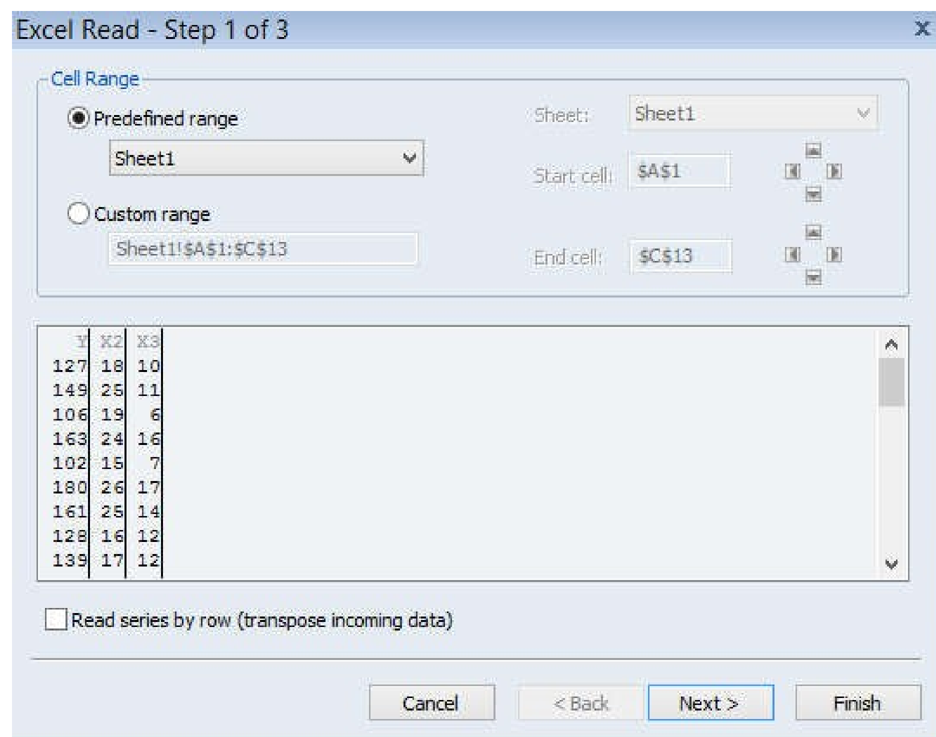


Hình 10



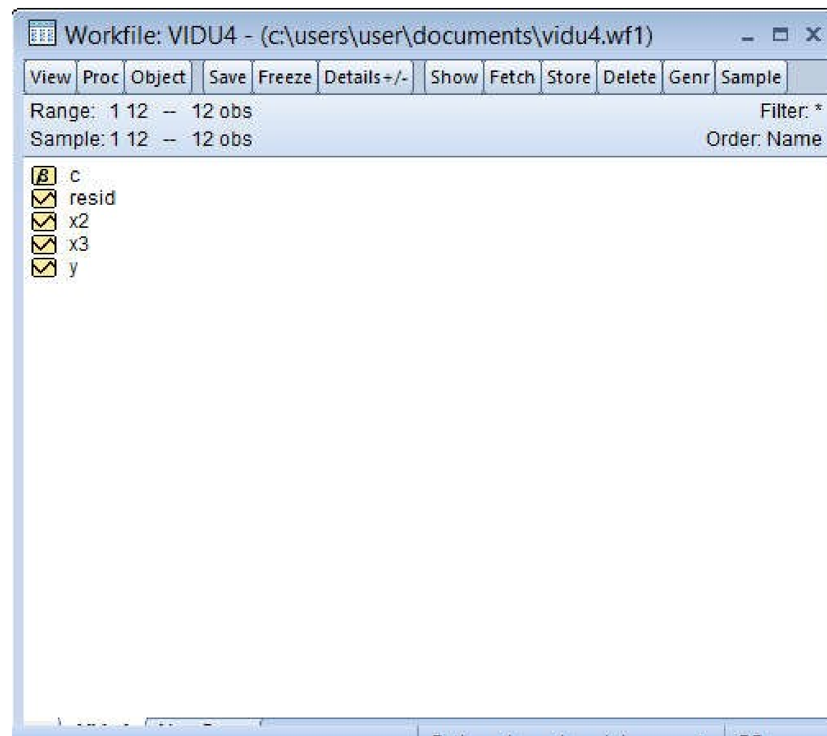
Hình 11

Chọn **Open** ta được kết quả như trong hình 12. Trong cửa sổ này chúng ta thấy có hai cột số liệu của X và Y tương ứng trong Sheet1 của File vidu 3.xls



Hình 12

Cuối cùng chọn **Finish** ta được kết quả như trong hình 13

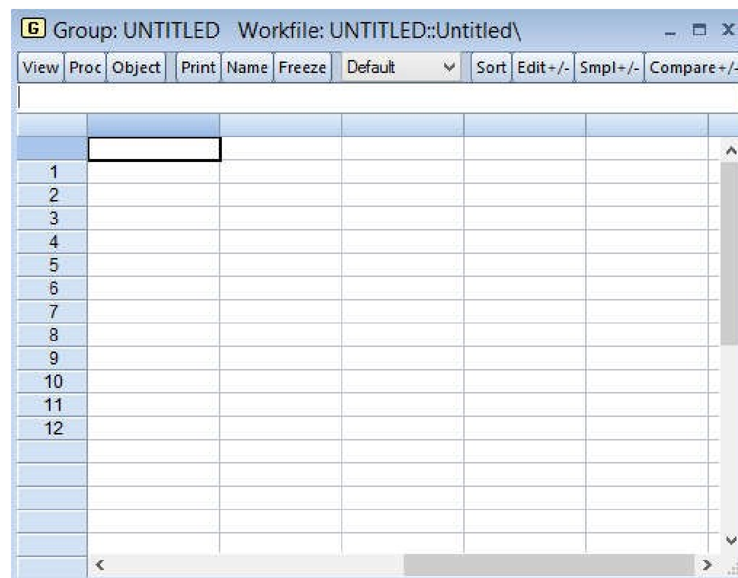


Hình 13

Lưu ý. Các bước trên được gọi là trích lọc dữ liệu từ một file dữ liệu có sẵn.

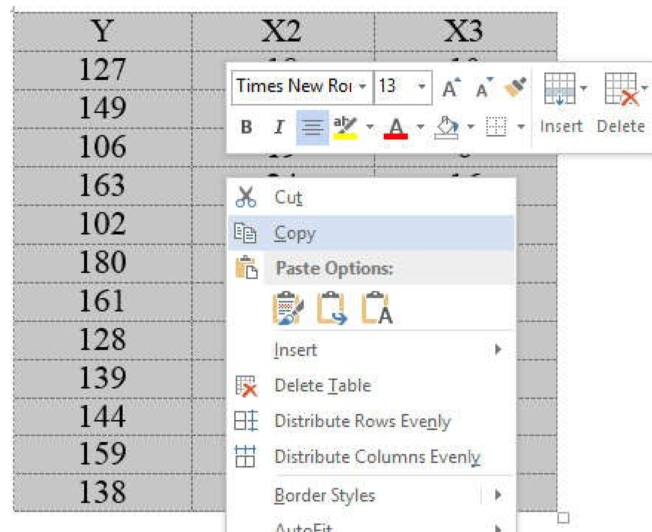
Ta có thể thực hiện copy trực tiếp từ một file Word hoặc Excel

Mở của số Group của Eview



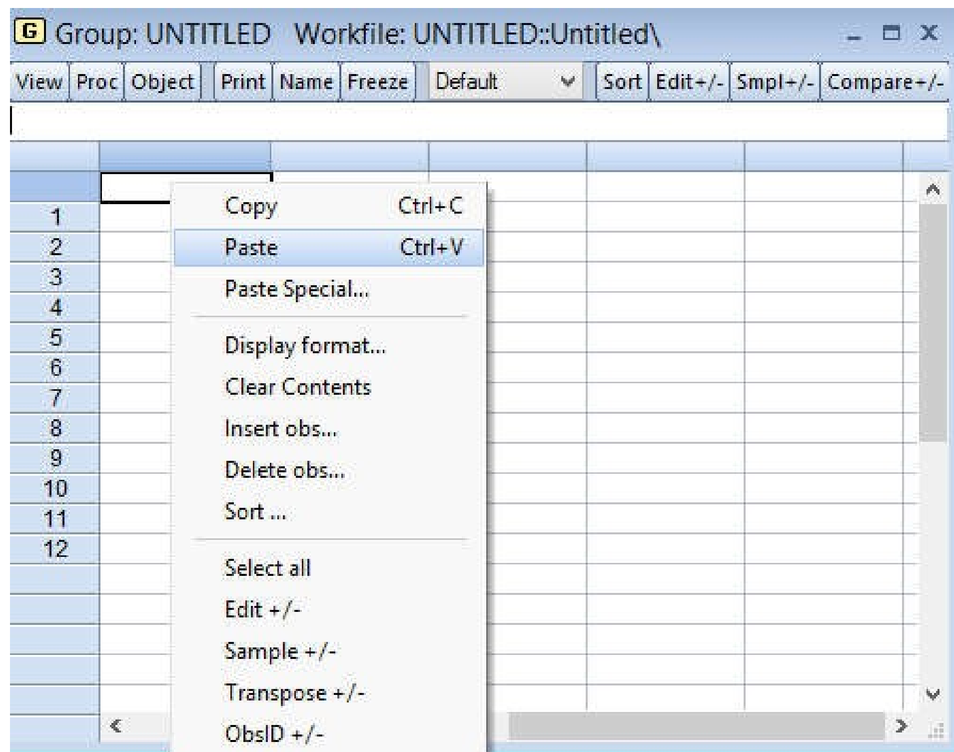
Hình 14

Từ file excel hoặc file word bôi đen rồi copy và paste vào file trên. Chẳng hạn ta có file word ta thực hiện như sau:



Hình 15

Ta paste vào của sổ Group như sau



Hình 16 Và

được kết quả như sau:

Group: UNTITLED   Workfile: UNTITLED::Untitled\

	Y	X2	X3
	Y	X2	X3
1	127.0000	18.00000	10.00000
2	149.0000	25.00000	11.00000
3	106.0000	19.00000	6.000000
4	163.0000	24.00000	16.00000
5	102.0000	15.00000	7.000000
6	180.0000	26.00000	17.00000
7	161.0000	25.00000	14.00000
8	128.0000	16.00000	12.00000
9	139.0000	17.00000	12.00000
10	144.0000	23.00000	12.00000
11	159.0000	22.00000	14.00000
12	138.0000	15.00000	15.00000

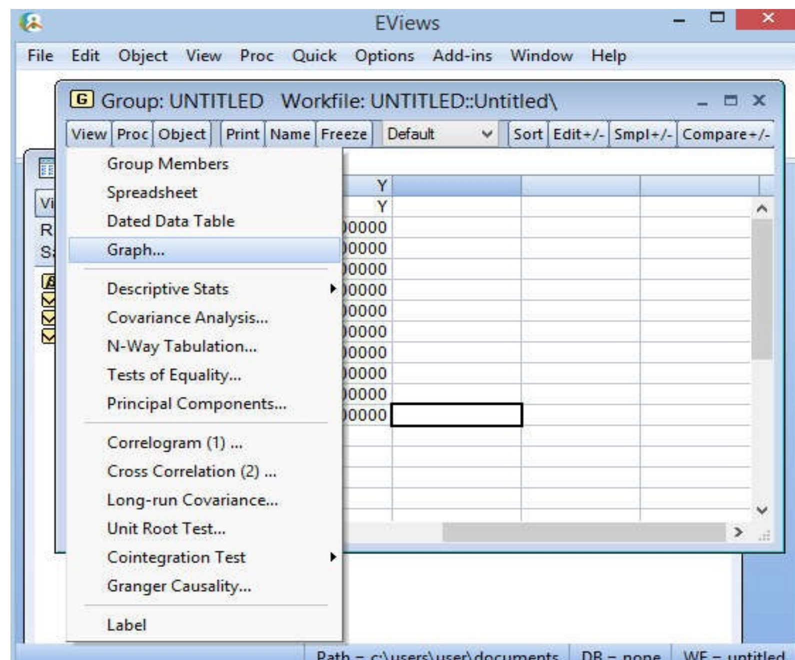
Hình 17

4. Vẽ đồ thị

4.1. Vẽ biểu đồ phân tán số liệu.

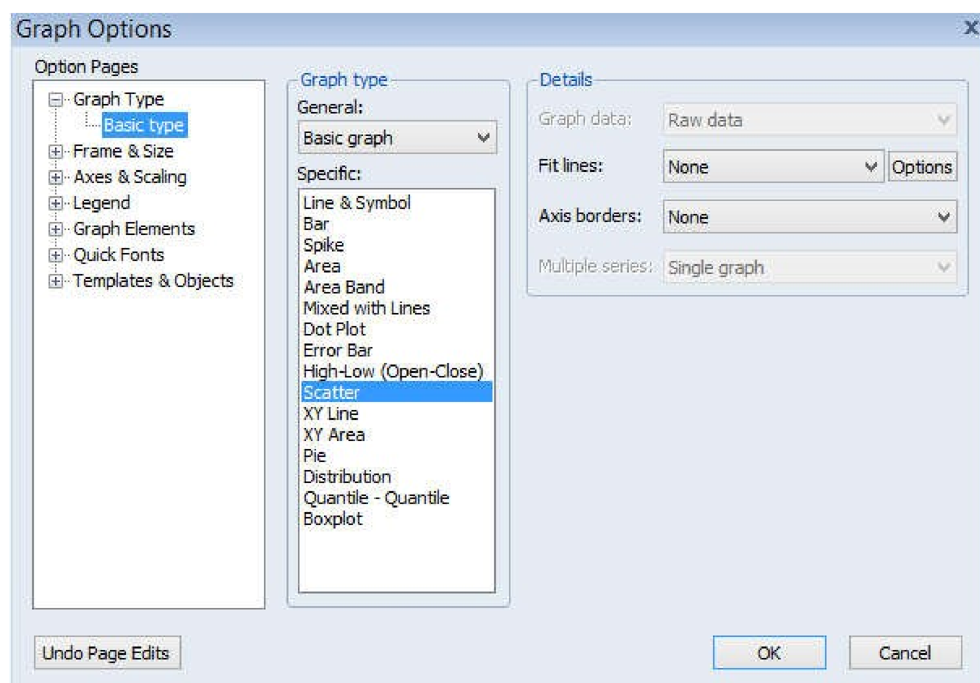
Mục đích của việc vẽ đồ thị này cho phép ta đánh giá sơ bộ về mối quan hệ cũng như hình dung được dạng hàm (mô hình) giữa hai biến với nhau. Để vẽ đồ thị phân tán của hai biến, chẳng hạn như trong ví dụ 3 ta vẽ đồ thị phân tán của Y và X.

Từ cửa sổ **Eviews** chọn **Quick→Graph**



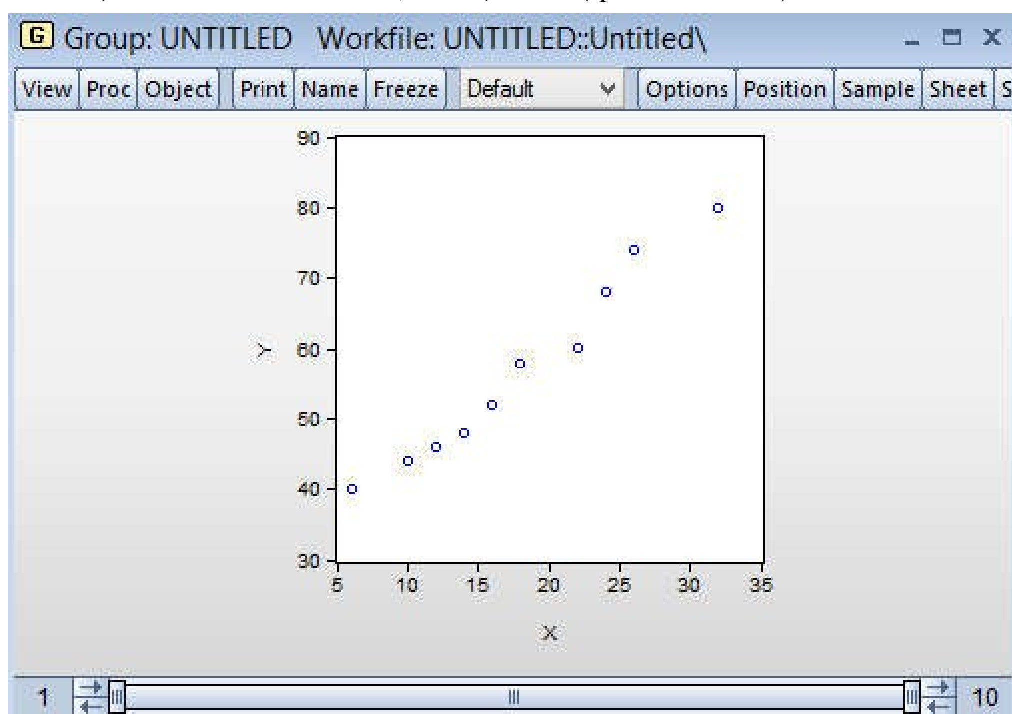
Hình 18

Ta được



Hình 19

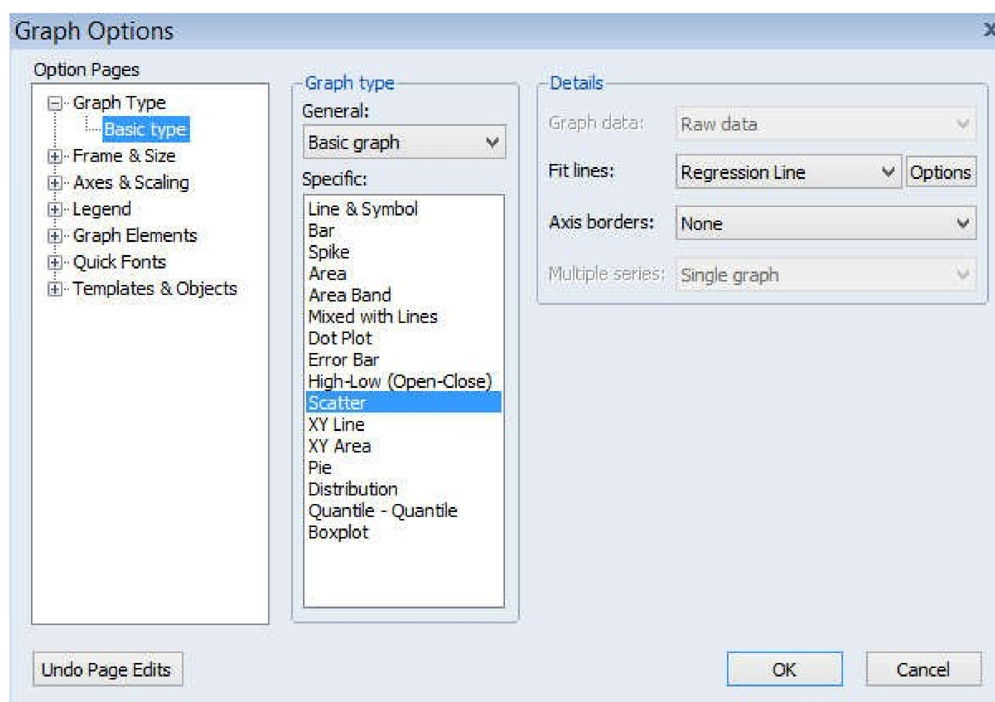
Ta chọn **Scatter** rồi nhấn **Ok**, ta được đồ thị phân tán dữ liệu như sau



Hình 20

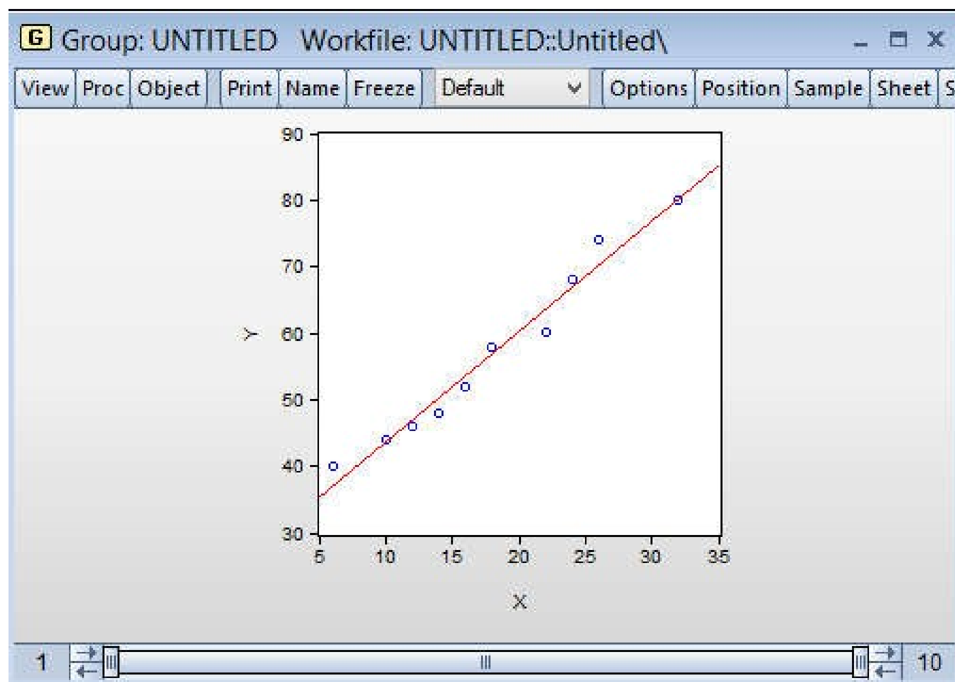
Làm tương tự như các bước trên ta có thể vẽ các loại đồ thị khác.

4.2. Vẽ đường hồi quy tuyến tính.



Hình 21

Thực hiện các bước tương tự như trên. Ta chọn **Scatter→Regression line** rồi nhấn **Ok**, ta được đồ thị đường hồi quy như sau:



Hình 22

5. Tìm hàm hồi quy tuyến tính mẫu (SRF)

Muốn tìm hàm hồi quy tuyến tính mẫu của Y theo X chẳng hạn như trong ví dụ 3 có nhiều cách làm sau đây tôi chỉ giới thiệu một cách đơn giản nhất.

Từ cửa sổ **Command** ta gõ dòng lệnh **ls y c x** và nhấn **Enter.** Ta có bảng hồi quy sau mà ta gọi là bảng **Equation**

EViews

File Edit Object View Proc Quick Options Add-ins Window Help

ls y c x

Equation: UNTITLED Workfile: UNTITLED::Untitled\

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: Y
Method: Least Squares
Date: 07/18/19 Time: 11:04
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	27.12500	1.979265	13.70458	0.0000
X	1.659722	0.101321	16.38082	0.0000

R-squared	0.971049	Mean dependent var	57.00000
Adjusted R-squared	0.967430	S.D. dependent var	13.47426
S.E. of regression	2.431706	Akaike info criterion	4.791920
Sum squared resid	47.30556	Schwarz criterion	4.852437
Log likelihood	-21.95960	Hannan-Quinn criter.	4.725533
F-statistic	268.3312	Durbin-Watson stat	1.783613
Prob(F-statistic)	0.000000		

Path = c:\users\user\documents DB = none WF = untitled

Hình 23

Các kết quả ở bảng trong **hình 23** lần lượt là

- ***Dependent Variable*** : Tên biến phụ thuộc
- ***Method: Least Squares*** : Phương pháp bình phương tối thiểu (nhỏ nhất).
- ***Date – Time*** : Ngày giờ thực hiện
- ***Sample*** : Số liệu mẫu 1 – 10
- ***Included observations*** : Cỡ mẫu là 10 (số các quan sát)
- Cột ***Variable*** : Các biến giải thích có trong mô hình (trong đó C là hệ số bị chặn)
- Cột ***Coefficient*** : Giá trị các hệ số hồ quy $\hat{\beta}_1; \hat{\beta}_2$. - Cột ***Std. Error*** : Sai số chuẩn của các hệ số hồi quy.

$$se(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)} \qquad se(\hat{\beta}_2) = \sqrt{var(\hat{\beta}_2)}$$

- Cột ***t – Statistic*** : Giá trị thống kê t tương ứng: $t_1 = \dfrac{\hat{\beta}_1}{se(\hat{\beta}_1)}; t_2 = \dfrac{\hat{\beta}_2}{se(\hat{\beta}_2)}$

(Trong đó t là đại lượng ngẫu nhiên có phân phối Student vớ bậc tự do (n – 2)).

- **Cột *Prob.*** : Giá trị xác suất (p – value) của thống kê t tương ứng

p_value$_1$ □ P t□ □ t ;p_value$_1$□$_2$ □ P t□ □ t_2□

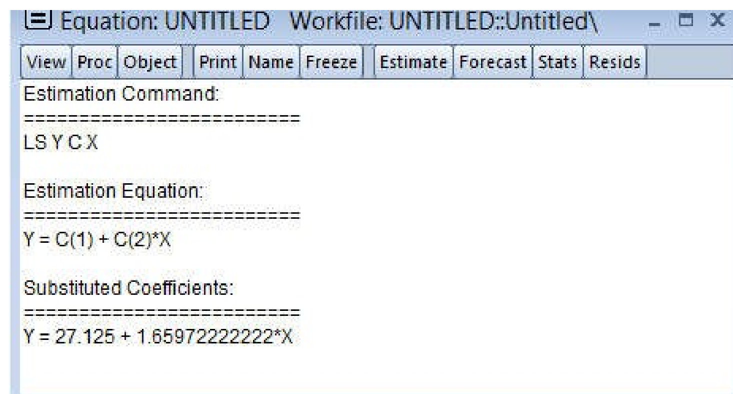
- ***R – Squared*** : Hệ số xác định mô hình (R^2)
- ***Adjusted R – Squared*** : Hệ số xác định có hiệu chỉnh ($\overline{R^2}$)
- ***S.E. of regression*** : Giá trị ước lượng cho σ : $\hat{□}$ (sai số chuẩn của hồi quy)

- ***Sum squared resid*** : Tổng bình phương các sai lệch (phần dư) (RSS)
- ***Log likelihood*** : Tiêu chuẩn ước lượng hợp lý (Logarit của hàm hợp lý)
- ***Durbin – Watson stat*** : Thống kê Durbin – Watson
- ***Mean dependent var*** : Giá trị trung bình mẫu của biến phụ thuộc
- ***S.D. dependent var*** : Độ lệch chuẩn mẫu của biến phụ thuộc
- ***Akaike info criterion*** : Tiêu chuẩn Akaike
- ***Schwarz info criterion*** : Tiêu chuẩn Schwarz
- ***F – Statistic*** : Giá trị của thống kê F
- ***Prob (F – Statistic)*** : Giá trị xác suất (p-value) của thống kê F tương ứng p_ value

□ P F□ □ F_statistic□

Với F là biến ngẫu nhiên có phân phố Fisher có bậc tự do (k − 1,n − k).

Muốn thể hiển đường hồi quy. Từ bảng **Equation→View→Representations**, ta có kết quả sau:



Hình 24

6. Một số hàm trong Eviews

LOG(X) : ln(X)

EXP(X) : e^X

ABS(X) : giá trị tuyệt đối của X

SQR(X) : căn bậc 2 của X

@SUM(X) : tổng của các X

@MEAN(X) : giá trị trung bình của X

@VAR(X) : phương sai của X

@COV(X,Y) : hiệp phương sai của X, Y

@COR(X,Y) : hệ số tương quan của X, Y

7. Cách tìm một số dạng hàm hồi quy

Giả sử ta có số liệu của các biến Y và X tại thời điểm t. Nếu tìm hàm hồi quy của Y_t theo X và Y_{t-1} (biến trễ thì câu lệnh sẽ là **y c x y(-1).**

Giả sử ta có số liệu của các biến Y và X. Nếu tìm hàm hồi quy của ln(Y) theo ln(X) thì câu lệnh sẽ là **log(y) c log(x).**

Giả sử ta có số liệu của các biến Y và X. Nếu tìm hàm hồi quy của Y theo $X\sqrt{}$thì câu lệnh sẽ là **y c sqr(x).**

Giả sử ta có số liệu của các biến Y và X. Nếu tìm hàm hồi quy của Y theo e^X thì câu lệnh sẽ là y **c exp(x).**

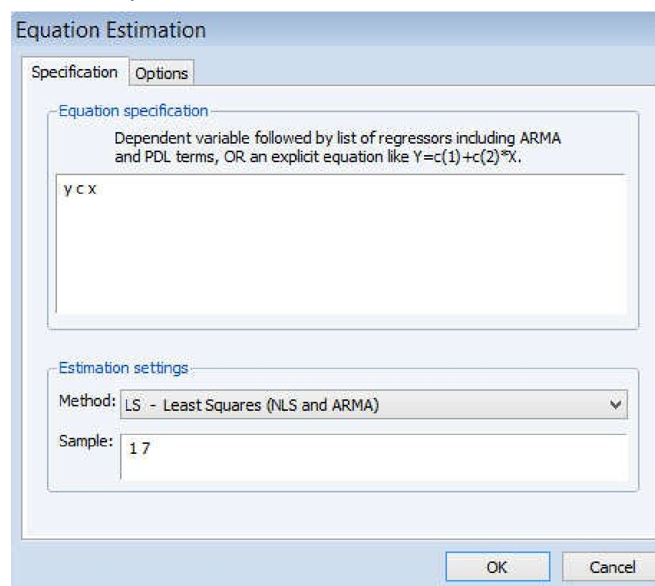
Giả sử ta có số liệu của các biến Y và X. Nếu tìm hàm hồi quy của Y theo X và X^2 thì câu lệnh sẽ là **y c x x^2.**

Giả sử ta có số liệu của các biến Y và X. Nếu tìm phương trình sai phân cấp 1 của Y theo X thì câu lệnh sẽ là d(**y) c d(x).**

Giả sử ta có số liệu của các biến Y và X. Nếu tìm phương trình sai phân cấp k của Y theo X thì câu lệnh sẽ là d(**y,k) c d(x,k).**

Nếu cần tìm hàm hồi quy nhưng không sử dụng hết các quan sát của mẫu, chẳng hạn ta tìm hàm hồi quy của Y theo X trong ví dụ 3 nhưng ta chỉ sử dụng 7 cặp quan sát đầu tiên. Khi đó ta thực hiện các thao tác như sau:

Từ bảng **Equation chọn Estimate,** ta có màn hình sau. Ta chỉnh 10 thành 7



Hình 25

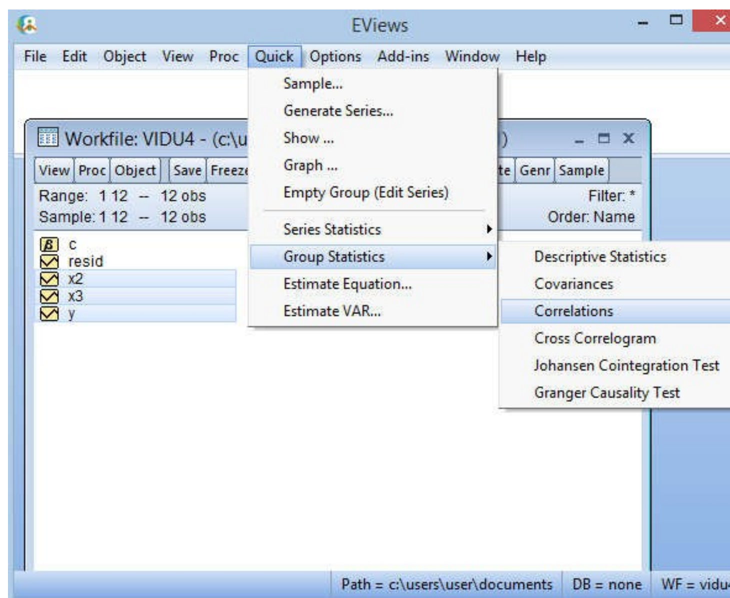
8. Tìm ma trận tương quan và ma trận hiệp phương sai của các hệ số hồi quy

8.1. Ma trận tương quan giữa các biến

Giả sử ta có mẫu gồm các biến Y, X2, X3 cho trong ví dụ 4. Để tìm ma trận tương quan của các biến này ta thực hiện như sau:

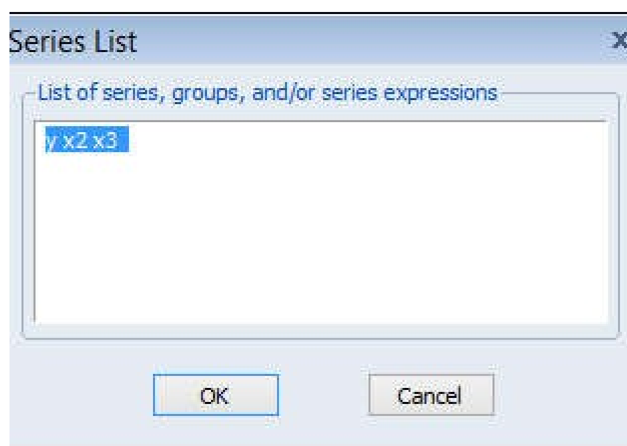
Từ cửa sổ **Eviews** chọn **Quick →Group Statistics →Correlations**.

Khi đó màn hình xuất hiện như sau:



Hình 26

Nhấp chuột sẽ xuất hiện cửa sổ sau



Hình 27

Sau đó **nhấn OK**, ta được ma trận tương quan như sau

Group: UNTITLED Workfile: VIDU4::Vidu4\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Correlation

	Y	X2	X3
Y	1.000000	0.782281	0.904627
X2	0.782281	1.000000	0.480173
X3	0.904627	0.480173	1.000000

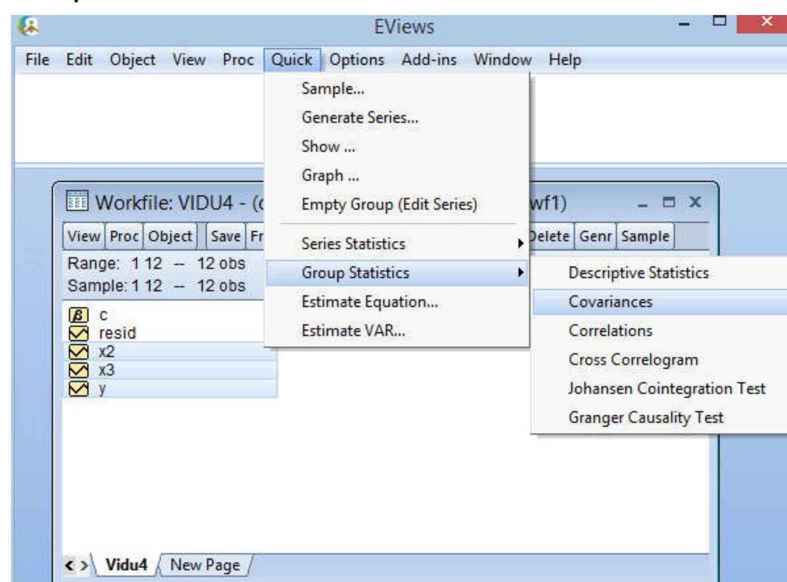
Hình 28

Ý nghĩa: Ma trận tương quan (**Correlation**) cho biết xu thế và mức độ tương quan tuyến tính giữa hai biến trong mô hình. Nhìn vào bảng ma trận tương quan ở trên ta thấy hệ số tương quan của X2 và X3 là 0.480173 khá nhỏ điều đó có nghĩa là X2 và X3 có tương quan tuyến tính ở mức độ yếu và tương quan thuận.

8.2. Ma trận hiệp phương sai giữa các biến

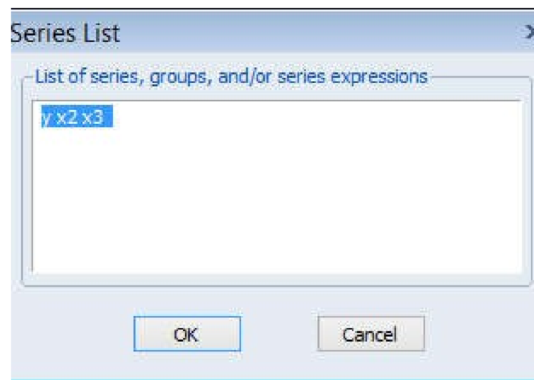
Giả sử ta có mẫu gồm các biến Y, X2, X3 cho trong ví dụ 4. Để tìm ma trận tương quan của các biến này ta thực hiện như sau:

Từ cửa sổ **Eviews** chọn **Quick →Group Statistics →Covariances** Khi đó màn hình xuất hiện như sau:



Hình 29

Nhấp chuột sẽ xuất hiện cửa sổ sau



Hình 30

Sau đó **nhấn OK**, ta được ma trận tương quan như sau

Group: UNTITLED Workfile: VIDU4::Vidu4\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Covariance

	Y	X2	X3
Y	493.7222	69.69444	64.52778
X2	69.69444	16.07639	6.180556
X3	64.52778	6.180556	10.30556

Hình 31

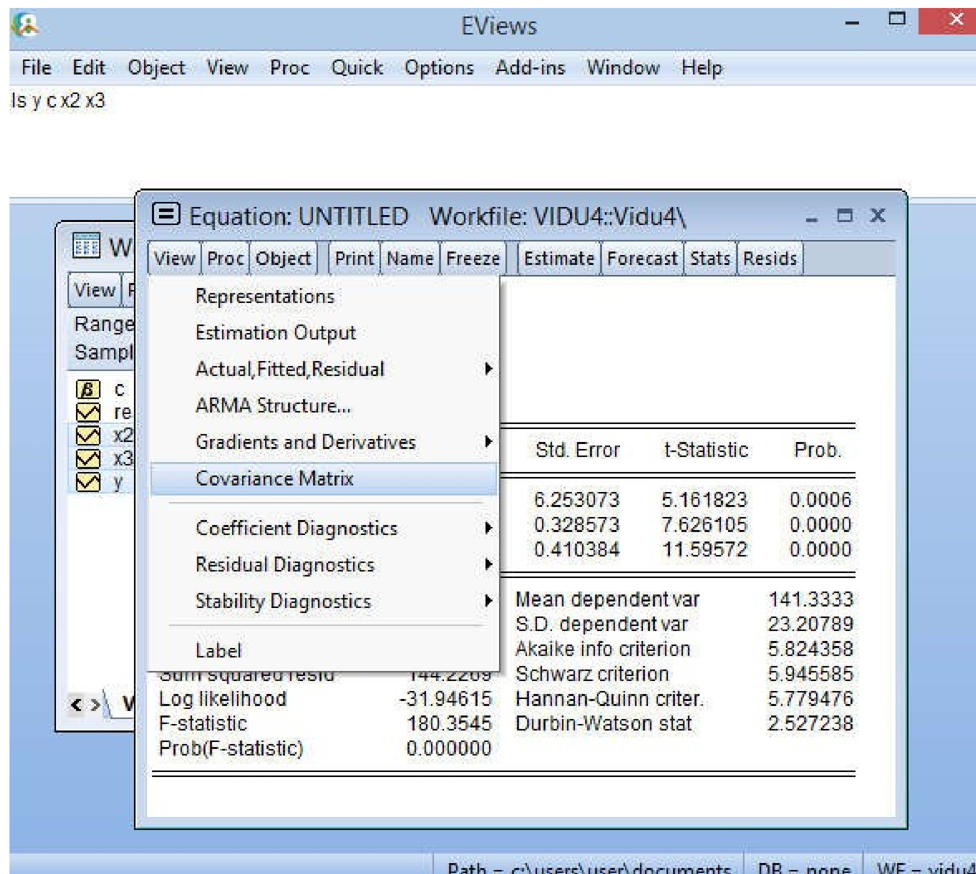
Ý nghĩa: Ma trận hiệp phương sai (Covariances) cho biết phương sai các biến nằm trên đường chéo chính.

8.3. Ma trận hiệp phương sai giữa các hệ số hồi quy

Giả sử ta có mẫu gồm các biến Y, X2, X3 cho trong ví dụ 4. Để tìm ma trận hiệp phương sai giữa các hệ số hồi quy, ta thực hiện như sau:

Từ cửa sổ **Equation** chọn **View →Covariance Matrix**.

Khi đó màn hình xuất hiện như sau:



Hình 32

Nhấp chuột, ta được ma trận hiệp phương sai giữa các hệ số hồi quy như sau

Equation: UNTITLED Workfile: VIDU4::Vidu4\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Coefficient Covariance Matrix

	C	X2	X3
C	39.10093	-1.416429	-0.727129
X2	-1.416429	0.107960	-0.064747
X3	-0.727129	-0.064747	0.168415

Hình 33

Ý nghĩa: Ma trận hiệp phương sai của các hệ số hồi quy (**Coefficient Covariance matrix**) cho biết phương sai các hệ số hồi quy nằm trên đường chéo chính, các thành phần còn lại là hiệp phương sai của những hệ số trong mô hình.

Chẳng hạn, ví dụ 4 bên trên. Nhìn vào ma trận hiệp phương sai bên trên ta có phương sai của các hệ số hồi quy là:

$$\operatorname{var}(\hat{\beta}_1) = 39.10093; \operatorname{var}(\hat{\beta}_2) = 0.107960; \operatorname{var}(\hat{\beta}_3) = 0.168415.$$

9. Bài toán tìm khoảng tin cậy cho các hệ số hồi quy (Khoảng tin cậy đối xứng)

Khoảng ước lượng các hệ số hồi quy tổng thể

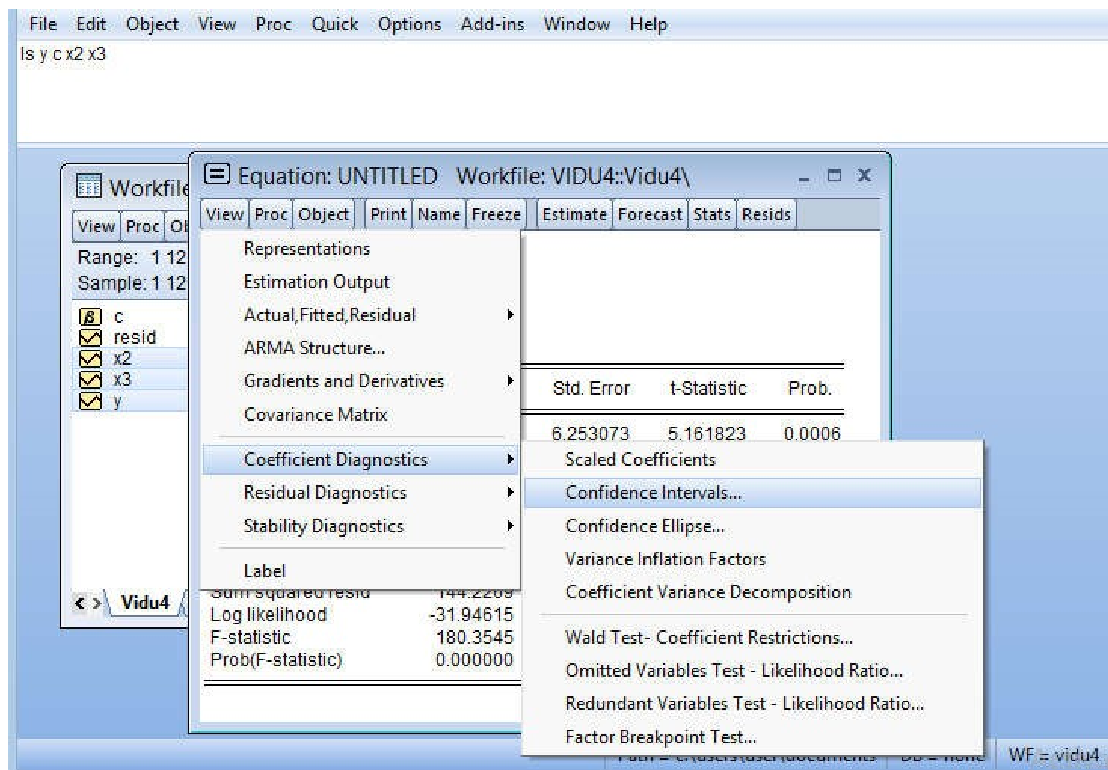
$$\beta_j \in \left(\hat{\beta}_j - C\,se(\hat{\beta}_j);\ \hat{\beta}_j + C\,se(\hat{\beta}_j)\right);\ j = 1,2,\ldots,k; \quad C = t^{n-k}_{\alpha/2}$$

Trong đó C là giá trị được dò trong bảng phân phối Student với bậc tự do là (n-k).

Giả sử ta có mẫu gồm các biến Y, X2, X3 cho trong ví dụ 4. Để tìm khoảng tin cậy cho các hệ số hồi quy tổng thể, ta thực hiện như sau:

Từ cửa sổ **Equation** chọn **View →Coefficient Diagnostics→confidence Intervals…**

Khi đó màn hình xuất hiện như sau:



Hình 34

Nhấp chuột, ta được kết quả sau

Equation: UNTITLED Workfile: VIDU4::Vidu4\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Coefficient Confidence Intervals
Date: 07/18/19 Time: 11:31
Sample: 1 12
Included observations: 12

Variable	Coefficient	90% CI Low	90% CI High	95% CI Low	95% CI High	99% CI Low	99% CI High
C	32.27726	20.81467	43.73985	18.13183	46.42270	11.95580	52.59872
X2	2.505729	1.903418	3.108040	1.762446	3.249012	1.437922	3.573536
X3	4.758693	4.006414	5.510973	3.830341	5.687045	3.425015	6.092372

Hình 35

Bảng trên là kết quả ước lượng khoảng tin cậy của các hệ số hồi quy tổng thể ứng với độ tin cậy 90%, 95% và 99%.

10. Bài toán dự báo

Khoảng dự báo giá trị trung bình

$$E(Y|X = X^0) \in \left(\hat{Y}_0 - C\,se(\hat{Y}_0);\hat{Y}_0 + C\,se(\hat{Y}_0)\right)$$

Khoảng dự báo giá trị cá biệt

$$Y_0 \in \left(\hat{Y}_0 - C\,se(Y_0 - \hat{Y}_0);\hat{Y}_0 + C\,se(Y_0 - \hat{Y}_0)\right)$$

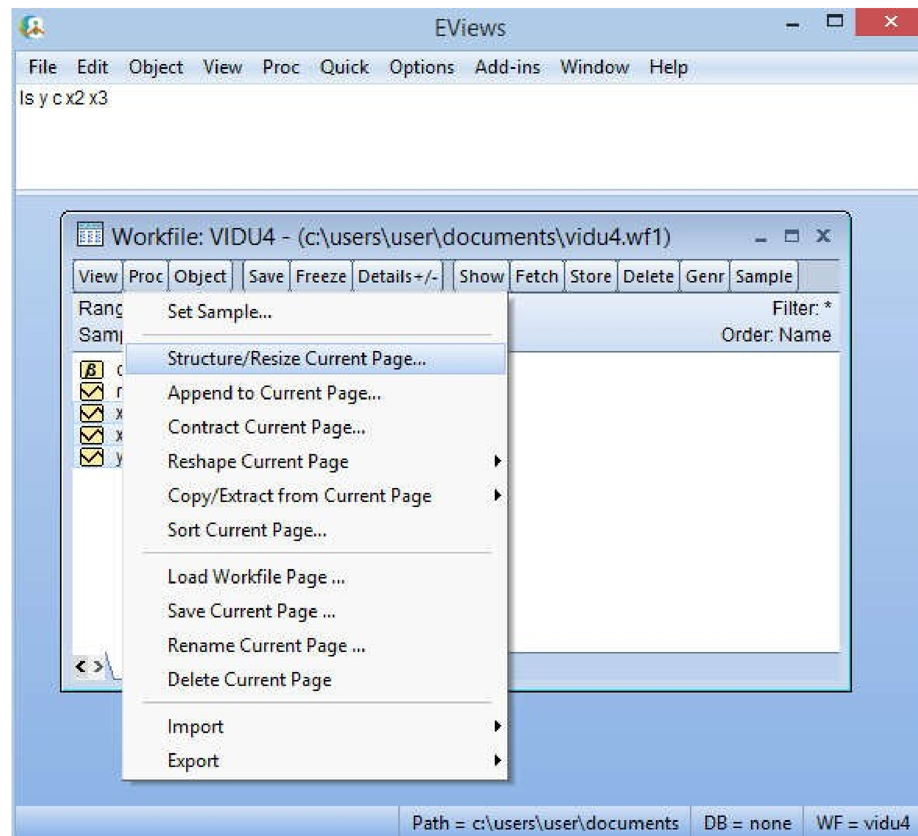
Đặt $Y_{DB} = \hat{Y}; Se1 = se(\hat{Y}_0); Se2 = se(Y_0 - \hat{Y}_0); C = @qtdist\left(1 - \frac{\alpha}{2}, n - k\right)$

MH → Equation

Xét ví dụ 4, để tìm khoảng dự báo giá trị trung bình và giá trị các biệt của Y khi $X_2 = 20, X_3 = 16$, với độ tin cậy 95%, ta thực hiện như sau:

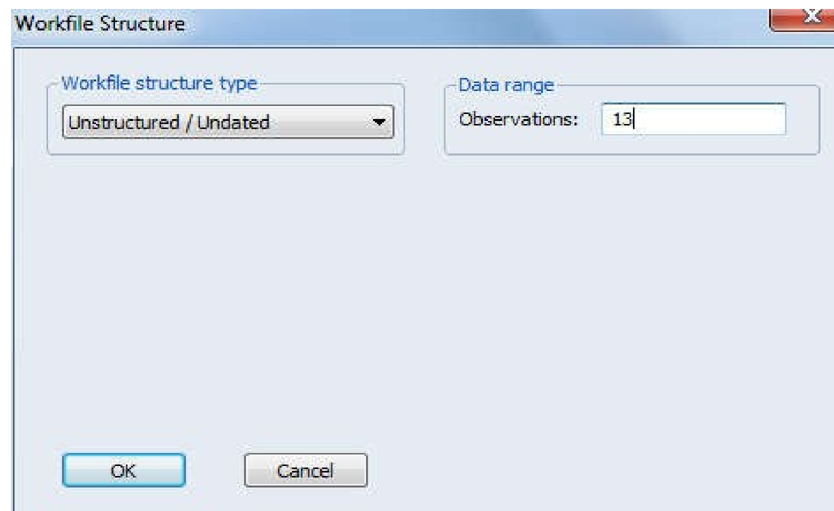
Bước 1. Nhập thêm dữ liệu vào bảng Group để dự báo

Từ bảng **Workfile**, chọn **Proc →Structure/Resize Current Page…**Màn hình sau



Hình 36

Nhấp chuột, màn hình sau xuất hiện. Ở ô quan sát (**Observations**) ta điều chỉnh 12 thành 13) như sau:



Hình 37

Nhấp OK. Từ bảng **Group.** Ta chọn **Edit+/- ,** sau đó nhập X_2 □ 20,X_3 □16 vào hàng số 13 có chữ NA như sau:

	X2	X3	Y
1	18.00000	10.00000	127.0000
2	25.00000	11.00000	149.0000
3	19.00000	6.000000	106.0000
4	24.00000	16.00000	163.0000
5	15.00000	7.000000	102.0000
6	26.00000	17.00000	180.0000
7	25.00000	14.00000	161.0000
8	16.00000	12.00000	128.0000
9	17.00000	12.00000	139.0000
10	23.00000	12.00000	144.0000
11	22.00000	14.00000	159.0000
12	15.00000	15.00000	138.0000
13	20.00000	16.00000	NA

Hình 38 Tắt cửa sổ **Group**.

Bước 2. Tính giá trị $\hat{Y}_0$ □ Y ;se Y_{DB} □ $_0$ □ $\hat{Y}_0$ □ □ se1;se Y □ $\hat{}_0$ □ □ se2. Từ bảng **Equation**. Chọn **forecast** màn hình xuất hiện như sau

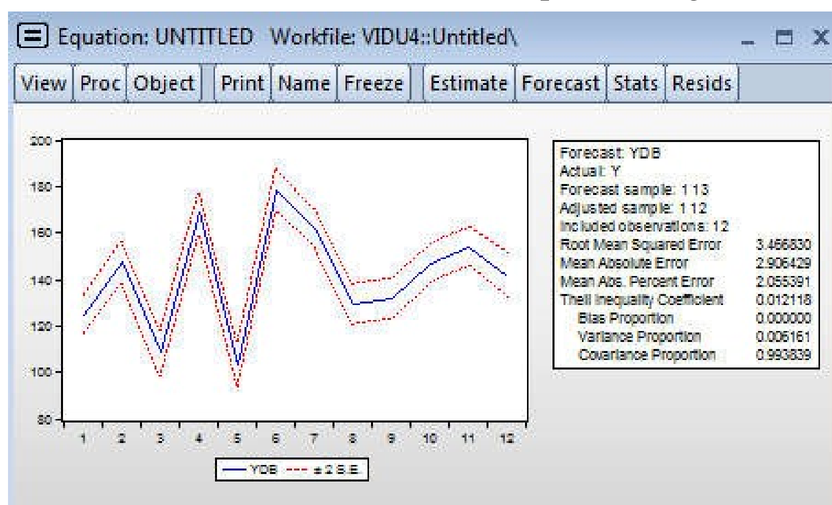
Equation: UNTITLED Workfile: VIDU4::Untitled\

Dependent Variable: Y
Method: Least Squares
Date: 06/07/13 Time: 21:31
Sample (adjusted): 1 12

Forecast
Forecast of — Equation: UNTITLED — Series: Y
Series names — Forecast name: YDB — S.E. (optional): Se1 — GARCH(optional):
Method — Static forecast (no dynamics in equation) — Structural (ignore ARMA) — Coef uncertainty in S.E. calc
Forecast sample — 1 13
Output — Forecast graph — Forecast evaluation
Insert actuals for out-of-sample observations
OK — Cancel

Hình 39

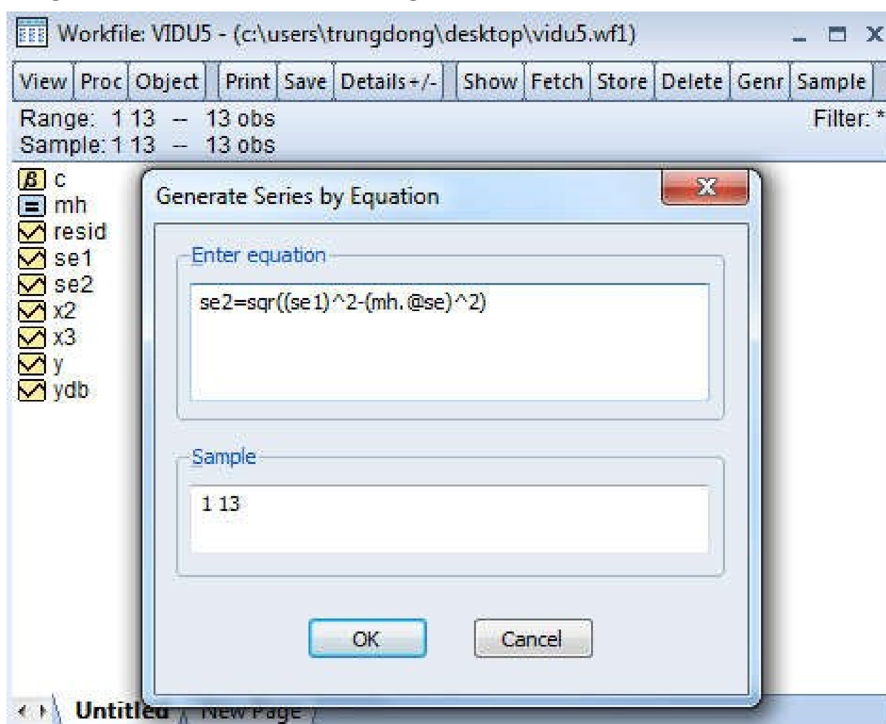
Ô **Forecast name** ta đổi Yf thành Y_{DB}, ô **S.E. (optional)** ta gõ Se1. **Nhấn OK.**



Hình 40

Tắt đồ thị dự báo

Từ bảng **Workfile**. Chọn **Genr** và gõ lệnh như sau rồi nhấn **Ok**.

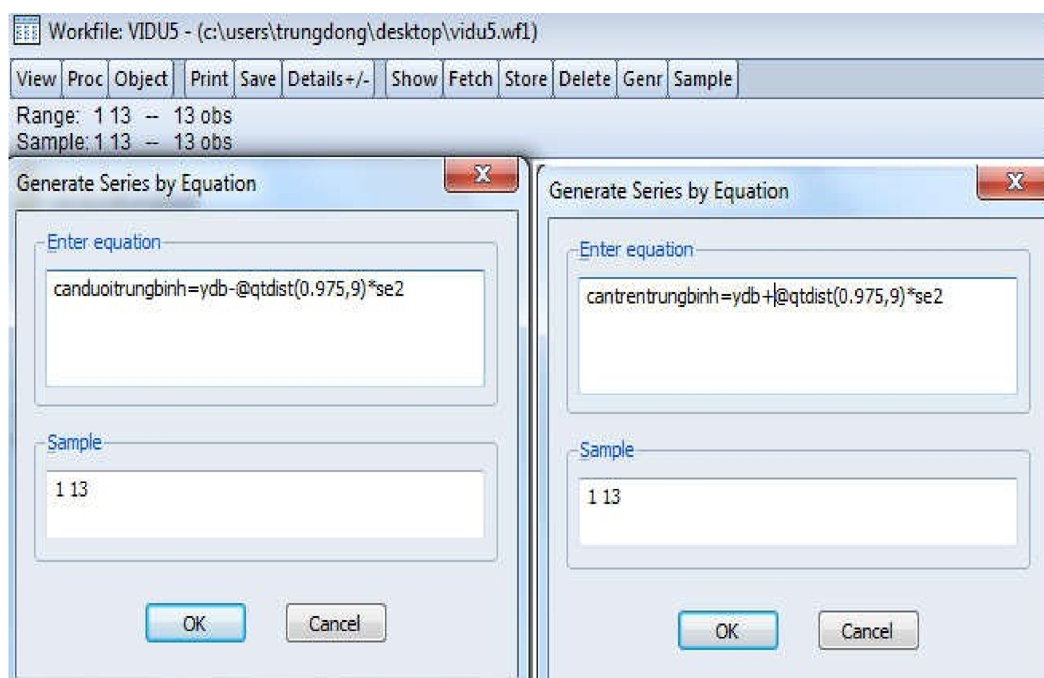


Hình 41

Bước 3. Tìm khoảng dự báo

- Dự báo giá trị trung bình

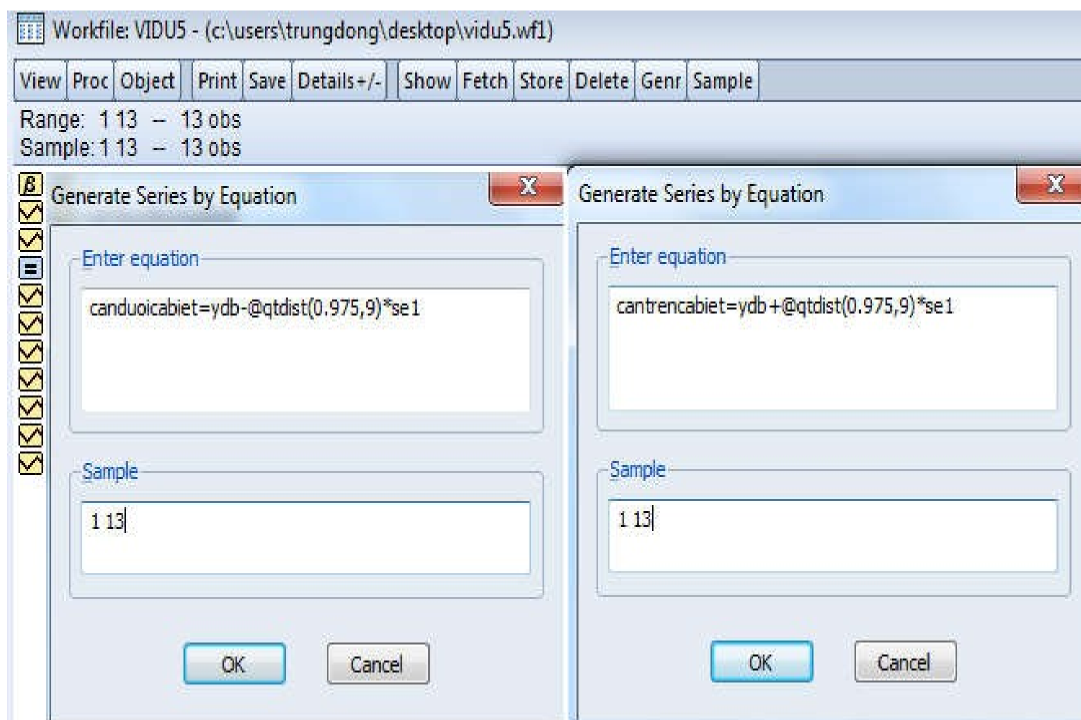
Từ bảng **Workfile**. Chọn **Genr** và gõ lệnh như sau rồi nhấn **Ok**.



Hình 42

Dự báo giá trị cá biệt.

Từ bảng **Workfile**. Chọn **Genr** và gõ lệnh như sau rồi nhấn **Ok**.



Hình 43

Để mở các kết quả trên cùng một bảng ta thực hiện như sau:

Từ của số **Workfile**, nhấn phím **Ctrl** rồi chọn canduoicabiet, cantrencabiet, canduoitrungbinh, cantrentrungbinh sau đó nhấn **Enter**, ta được kết quả sau (lưu ý nhìn vào hàng thứ 13)

Group: UNTITLED Workfile: VIDU5::Untitled\

View | Proc | Object | Print | Name | Freeze | Default | Sort | Transpose | Edit+/- | Smpl+/- | Title | Sample

obs	CANDUOITRUNGBINH	CANTRENTRUNGBINH	CANDUOICABIET	CANTRENCABIET
obs	CANDUOITRUNGBINH	CANTRENTRUNGBINH	CANDUOICABIET	CANTRENCABIET
1	121.7060	128.2287	115.3422	134.5925
2	142.4540	152.0782	137.0112	157.5210
3	102.5283	114.3482	97.62467	119.2519
4	164.3718	172.7359	158.5791	178.5286
5	97.92959	108.4185	92.70930	113.6388
6	173.1950	183.4530	167.9166	188.7314
7	157.5713	165.5131	151.6541	171.4303
8	125.3324	133.6141	119.5157	139.4308
9	128.3832	135.5748	122.2354	141.7225
10	143.7224	150.3043	137.3782	156.6485
11	150.9932	157.0569	144.4752	163.5748
12	134.9086	147.5786	130.1920	152.2952
13	153.9864	163.0754	148.3989	168.6630

Hình 44

Vậy khoảng dự báo giá trị trung bình và giá trị cá biệt của Y là

CANDUOITB	CANTRENTB	CANDUOICB	CANTRENCB
153.9864	163.0754	148.3989	168.6630

11. Định mẫu

Trước hết ta xét ví dụ sau

Ví dụ 5. Bảng số liệu sau cho biết số liệu về lượng hàng bán được (Y tấn/tháng), giá bán (X ngàn đồng/kg) ở 20 khu vực bán và được khảo sát tại hai nơi là Thành phố và Nông thôn.

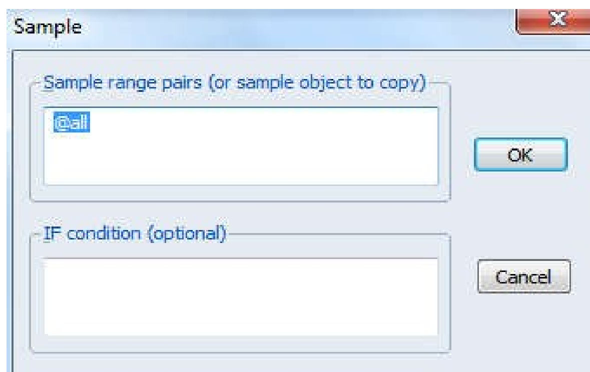
STT	Y	X	Z	STT	Y	X	Z
1	20	2	1	11	14	5	0
2	19	3	0	12	14	6	1
3	18	3	1	13	13	6	0
4	18	4	0	14	12	7	1
5	17	4	1	15	12	7	0
6	17	3	1	16	15	5	1
7	16	4	0	17	16	4	0
8	16	4	1	18	12	7	1
9	15	5	1	19	10	8	0
10	15	5	1	20	11	8	1

Bảng 8

Trong đó Z là biến giả với Z = 0 : khảo sát ở nông thôn; Z = 1 : khảo sát ở thành thị

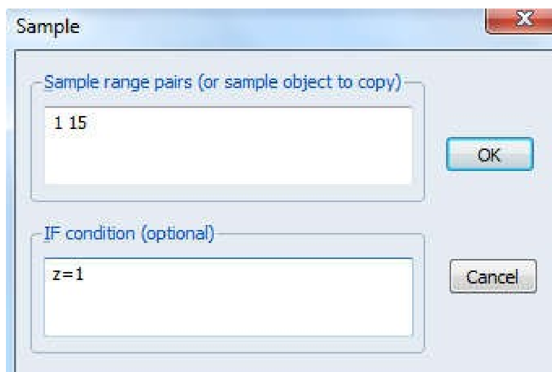
Có nhiều trường hợp ta không sử dụng hết các số liệu của mẫu ban đầu, hay chỉ cần khảo sát sự phụ thuộc khi biến giả nhận một giá trị nào đó.

Để định mẫu lại, từ cửa sổ **Workfile** chọn **Sample**, màn hình xuất hiện như



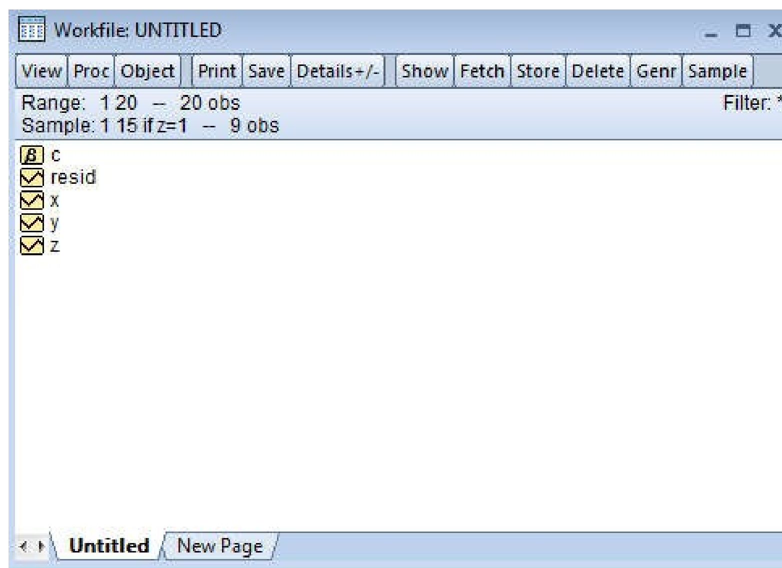
Hình 45

Chẳng hạn ta chỉ khảo sát 15 mẫu đầu tiên và ở khu vực Thành phố ứng với Z = 1. Ta khai báo vào ô Sample range pairs và IF condition nhưtrong hình sau



Hình 46

Nhấn OK, ta thấy có sự thay đổi trong cửa sổ Workfile như sau

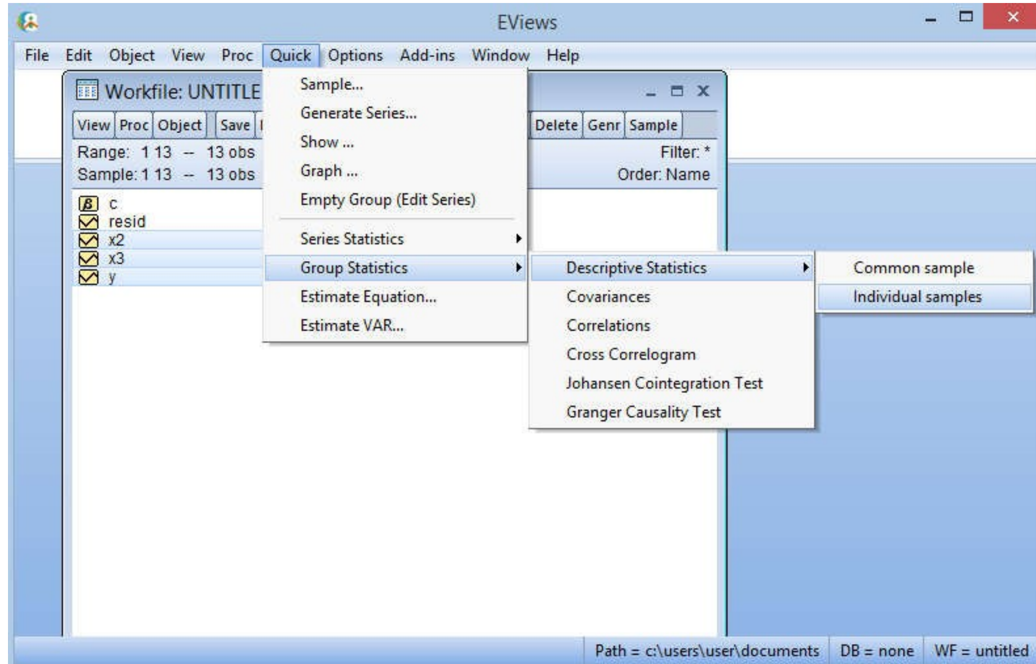


Hình 47

12. Tính các giá trị thống kê

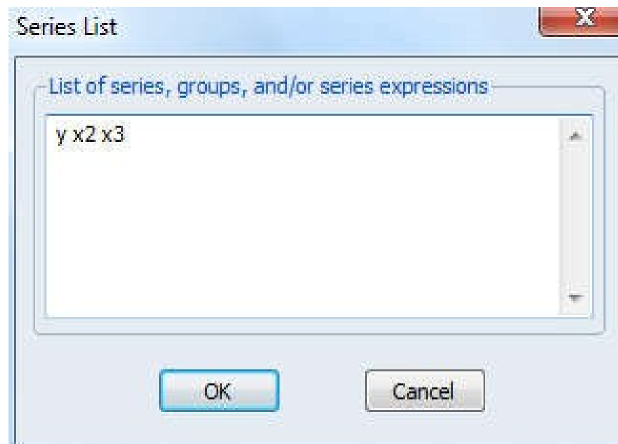
Để tính các giá trị thống kê như Trung bình, trung vị, độ lệch chuẩn, …của các biến có trong mô hình chẳng hạn vớ số liệu cho trong ví dụ 4 ta làm như sau:

Từ cửa sổ **EViews** chọn **Quick →Group Statistics →Descriptive statistics →Common sample**, như hình sau



Hình 48

Nhấp chuột và nhập tên các biến vào cửa sổ **Series List** như hình sau



Hình 49

Nhấp OK, ta được bảng các giá trị thống kê sau:

Group: UNTITLED Workfile: UNTITLED::Untitled\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

	Y	X2	X3
Mean	141.3333	20.41667	12.16667
Median	141.5000	20.50000	12.00000
Maximum	180.0000	26.00000	17.00000
Minimum	102.0000	15.00000	6.000000
Std. Dev.	23.20789	4.187825	3.352972
Skewness	-0.218444	-0.034924	-0.453123
Kurtosis	2.301560	1.401529	2.378514
Jarque-Bera	0.339345	1.279994	0.603763
Probability	0.843941	0.527294	0.739426
Sum	1696.000	245.0000	146.0000
Sum Sq. Dev.	5924.667	192.9167	123.6667
Observations	12	12	12

Hình 50 Giải

thích :

-Mean : trung bình.

-Median : trung vị

-Maximum : Giá trị lớn nhất - Minimum : Giá trị nhỏ nhất

-Std. Dev : Độ lệch chuẩn

-Skewness : Hệ số bất đối xứng

-Kurtosis : Hệ số nhọn

-Jarque – Bera : Kiểm định phân phối chuẩn

-Sum : Tổng các quan sát

-Sum sq. Dev : Độ lệch chuẩn của tổng bình phương

-Observations : Số quan sát (cỡ mẫu)

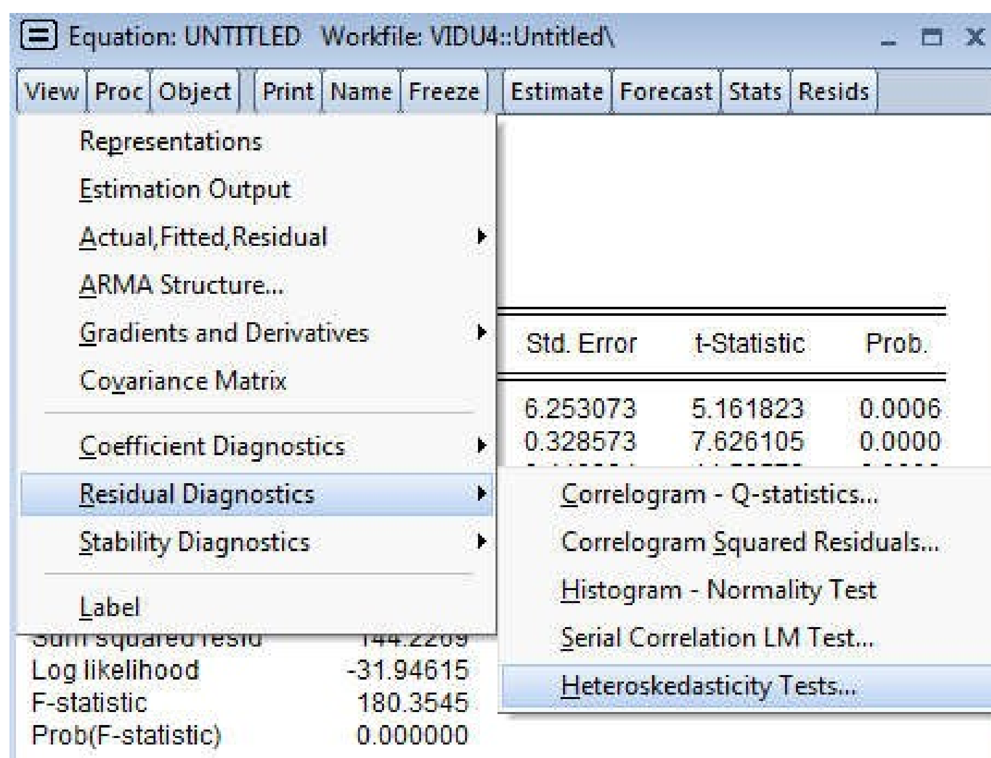
13. Các bài toán kiểm định giả thiết mô hình

13.1. Kiểm định phương sai thay đổi

13.1.1. Kiểm định White

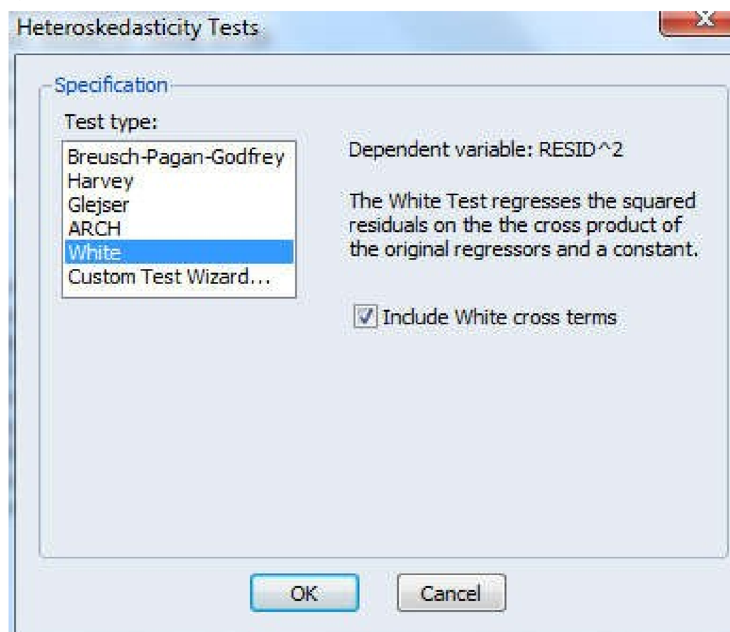
Chẳng hạn như trong ví dụ 4.

Để thực hiện việc kiểm định **White** bằng **Eview**, sau khi ước lượng mô hình hồi quy mẫu, từ cửa sổ **Equation** chọn **View→Residual Diagnostics → Heteroskedasticity tests…** Khi đó màn hình sẽ như sau:



Hình 51

Nhấp chuột, màn hình như sau



Hình 52

Ta chọn **White**, rồi **nhấn Ok**. Ta có kết quả như sau:

Equation: UNTITLED Workfile: VIDU4::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Heteroskedasticity Test: White

F-statistic	0.843820	Prob. F(5,6)	0.5645
Obs*R-squared	4.954369	Prob. Chi-Square(5)	0.4215
Scaled explained SS	2.033496	Prob. Chi-Square(5)	0.8445

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 06/08/13 Time: 21:40
Sample: 1 12
Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-302.2919	201.0729	-1.503394	0.1834
X2	31.71005	20.36090	1.557400	0.1704
X2^2	-0.879882	0.520465	-1.690568	0.1419
X2*X3	0.206625	0.536312	0.385270	0.7133
X3	0.587451	10.18193	0.057695	0.9559
X3^2	-0.075395	0.534322	-0.141104	0.8924

R-squared	0.412864	Mean dependent var	12.01891
Adjusted R-squared	-0.076416	S.D. dependent var	15.16494
S.E. of regression	15.73369	Akaike info criterion	8.656339
Sum squared resid	1485.295	Schwarz criterion	8.898792
Log likelihood	-45.93803	Hannan-Quinn criter.	8.566574
F-statistic	0.843820	Durbin-Watson stat	2.579780
Prob(F-statistic)	0.564521		

Hình 53

Ta đặt bài toán kiểm định như sau: H_0: Mô hình không xảy ra hiện tượng phương sai thay đổi; H_1: Mô hình xảy ra hiện tượng phương sai thay đổi.

Từ bảng kiểm định White ở trên, ta có P_value □ 0.4215 □□ cho trước nên chấp nhận H_0. Vậy mô hình không xảy ra hiện tượng phương sai thay đổi.

13.1.2. Kiểm định Glejser.

Ta thực hiện các bước như trong kiểm định **White** nhưng ta chọn **Glejser**, rồi **nhấn Ok**. Ta có kết quả như sau:

Equation: UNTITLED Workfile: VIDU4::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Heteroskedasticity Test: Glejser

F-statistic	0.651938	Prob. F(2,9)	0.5440
Obs*R-squared	1.518508	Prob. Chi-Square(2)	0.4680
Scaled explained SS	0.931349	Prob. Chi-Square(2)	0.6277

Test Equation:
Dependent Variable: ARESID
Method: Least Squares
Date: 06/08/13 Time: 21:42
Sample: 1 12
Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.690494	3.185754	0.844539	0.4203
X2	-0.127881	0.167398	-0.763934	0.4645
X3	0.232343	0.209078	1.111273	0.2953

R-squared	0.126542	Mean dependent var	2.906429
Adjusted R-squared	-0.067559	S.D. dependent var	1.973897
S.E. of regression	2.039485	Akaike info criterion	4.475590
Sum squared resid	37.43551	Schwarz criterion	4.596817
Log likelihood	-23.85354	Hannan-Quinn criter.	4.430708
F-statistic	0.651938	Durbin-Watson stat	1.830671
Prob(F-statistic)	0.543987		

Hình 54

Ta đặt bài toán kiểm định như sau: H :$_0$ Mô hình không xảy ra hiện tượng phương sai thay đổi; H :$_1$ Mô hình xảy ra hiện tượng phương sai thay đổi.

Từ bảng kiểm định **Glejser** ở trên, ta có P_value 0.4680 cho trước nên chấp nhận H .$_0$ Vậy mô hình không xảy ra hiện tượng phương sai thay đổi.

13.1.3. Kiểm định Breusch-Pagan-Godfrey

Ta thực hiện các bước tương tự như kiểm định **White** nhưng ta chọn **BreuschPagan-Godfrey**, rồi **nhấn Ok**. Ta có kết quả như sau:

Equation: UNTITLED   Workfile: VIDU4::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	0.596913	Prob. F(2,9)	0.5709
Obs*R-squared	1.405352	Prob. Chi-Square(2)	0.4953
Scaled explained SS	0.576820	Prob. Chi-Square(2)	0.7495

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 06/08/13   Time: 21:49
Sample: 1 12
Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.63993	24.60707	0.432393	0.6756
X2	-0.953765	1.292998	-0.737639	0.4795
X3	1.713837	1.614940	1.061239	0.3162

R-squared	0.117113	Mean dependent var	12.01891
Adjusted R-squared	-0.079085	S.D. dependent var	15.16494
S.E. of regression	15.75319	Akaike info criterion	8.564280
Sum squared resid	2233.466	Schwarz criterion	8.685507
Log likelihood	-48.38568	Hannan-Quinn criter.	8.519398
F-statistic	0.596913	Durbin-Watson stat	2.193097
Prob(F-statistic)	0.570918		

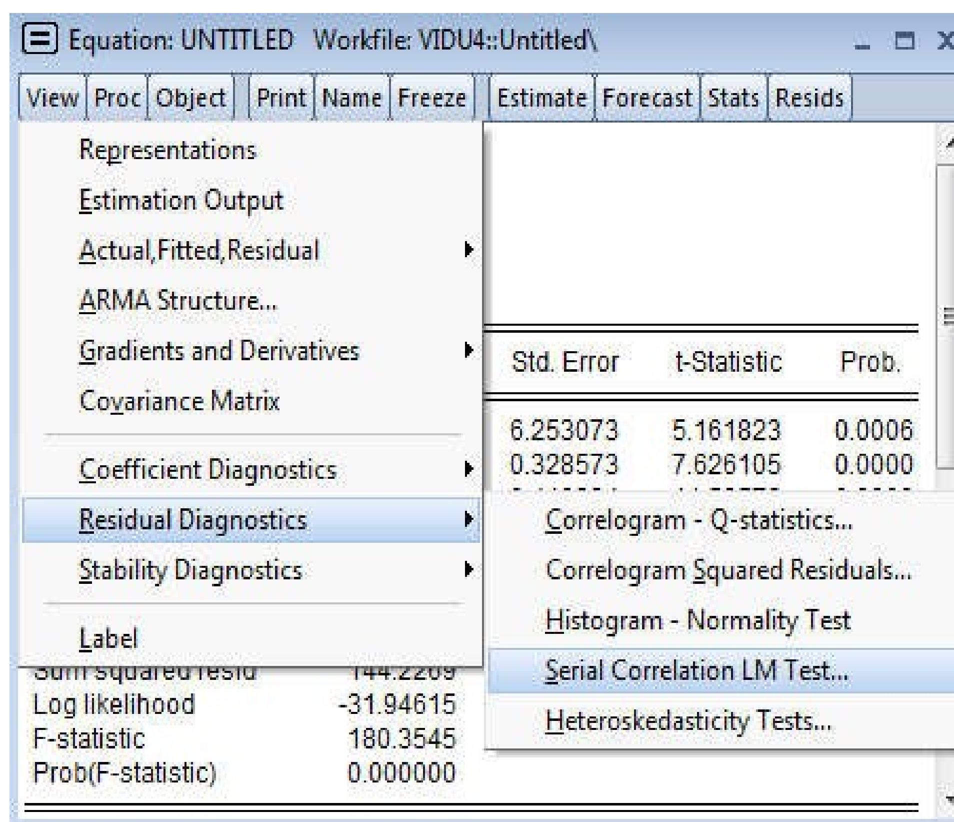
Hình 55

Ta đặt bài toán kiểm định như sau: H_0: Mô hình không xảy ra hiện tượng phương sai thay đổi; H_1: Mô hình xảy ra hiện tượng phương sai thay đổi.

Từ bảng kiểm định **Breusch – Pagan - Godfrey** ở trên, ta có P_ value □ 0.4953 □□ cho trước nên chấp nhận H_0. Vậy mô hình không xảy ra hiện tượng phương sai thay đổi.

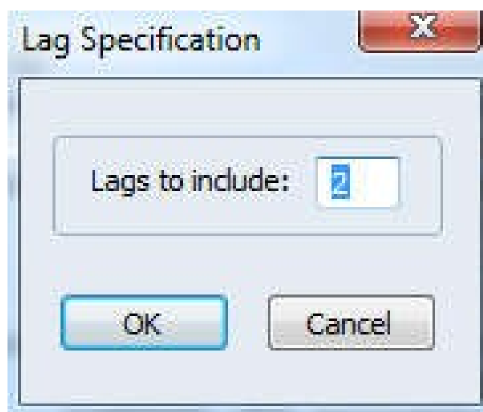
13.2. Kiểm định tự tương quan (Kiểm định BG) Chẳng hạn như trong ví dụ 4.

Để thực hiện việc kiểm định **BG** bằng Eview, sau khi ước lượng mô hình hồi quy mẫu, từ cửa sổ **Equation** chọn **View→Residual Diagnostics → Serial Correlation LM test…** Khi đó màn hình sẽ xuất hiện như sau:



Hình 56

Nhấp chuột, cửa sổ sau xuất hiện như sau:



Hình 57

Ô **Lags to indude** ta gõ bậc tự tương quan vào (ví dụ như tự tương quan là bậc 2) **Nhấn Ok**. Ta có kết quả như sau:

Equation: UNTITLED Workfile: VIDU4::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	0.481309	Prob. F(2,7)	0.6370
Obs*R-squared	1.450706	Prob. Chi-Square(2)	0.4842

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 06/08/13 Time: 22:00
Sample: 1 12
Included observations: 12
Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.843019	6.704450	-0.125740	0.9035
X2	0.027662	0.353556	0.078238	0.9398
X3	0.031521	0.442767	0.071191	0.9452
RESID(-1)	-0.321367	0.384510	-0.835784	0.4309
RESID(-2)	0.127392	0.426641	0.298592	0.7739

R-squared	0.120892	Mean dependent var	2.04E-14
Adjusted R-squared	-0.381455	S.D. dependent var	3.620986
S.E. of regression	4.255938	Akaike info criterion	6.028844
Sum squared resid	126.7910	Schwarz criterion	6.230888
Log likelihood	-31.17306	Hannan-Quinn criter.	5.954040
F-statistic	0.240655	Durbin-Watson stat	1.891410
Prob(F-statistic)	0.906546		

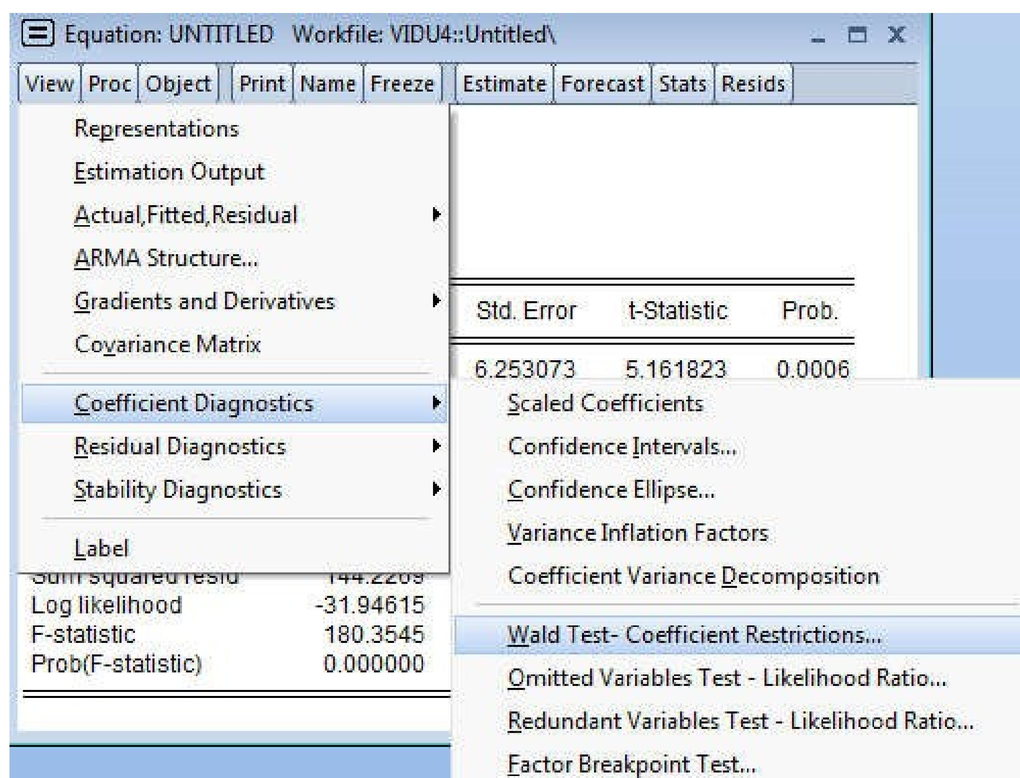
Hình 58

Ta đặt bài toán kiểm định như sau: H_0: Mô hình không xảy ra hiện tượng tự tương quan bậc 2; H_1: Mô hình xảy ra hiện tượng tự tương quan bậc 2.

Từ bảng kiểm định BG ở trên, ta có P_ value □ 0.4842 □□ cho trước nên chấp nhận H_0. Vậy mô hình không xảy ra hiện tượng tự tương quan bậc 2.

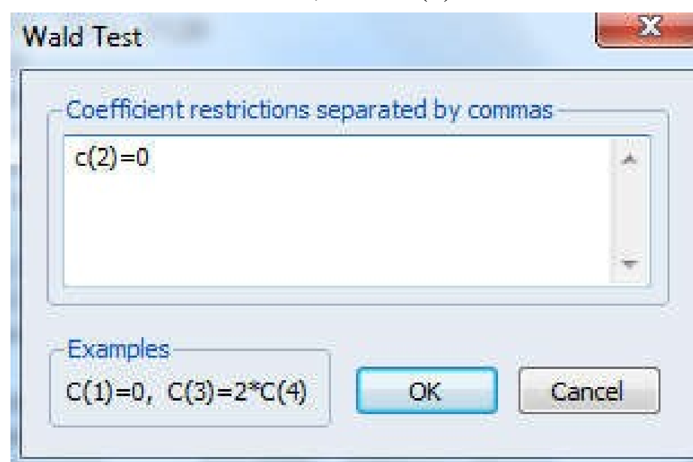
13.3. Kiểm định biến có cần thiết trong mô hình hay không (Kiểm định Wald) Chẳng hạn như trong ví dụ 4.

Để thực hiện việc kiểm định **Wald** bằng **Eview**, sau khi ước lượng mô hình hồi quy mẫu, từ cửa sổ **Equation** chọn **View→Coefficient Diagnostics → Wald test – Coefficient Restrictions…** Khi đó màn hình sẽ như sau:



Hình 59

Nhấp chuột ta có cửa sổ sau xuất hiện: Gõ c(2)=0 vào



Hình 60 **Nhấp**

Ok. Ta được kết quả như sau:

Equation: UNTITLED   Workfile: VIDU4::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
t-statistic	7.626105	9	0.0000
F-statistic	58.15747	(1, 9)	0.0000
Chi-square	58.15747	1	0.0000

Null Hypothesis: C(2)=0
Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(2)	2.505729	0.328573

Restrictions are linear in coefficients.

Hình 61

Ta đặt bài toán kiểm định như sau: H_0: Biến X2 không cần thiết trong mô hình;

H_1: Biến X2 cần thiết trong mô hình.

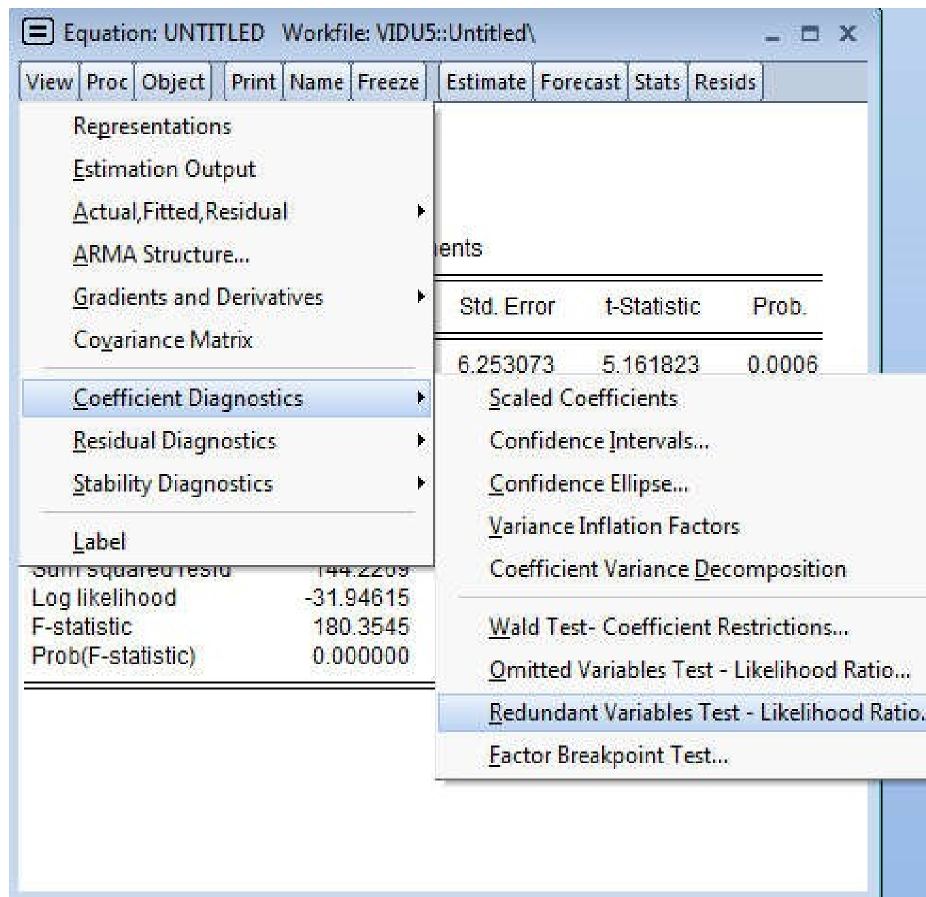
Từ bảng kiểm định **Wald** ở trên, ta có P_value □ 0.0000 □□ cho trước nên bác bỏ H_0. Vậy X2 cần thiết trong mô hình.

Lưu ý: Trong trường hợp này ta chỉ khảo sát X2 nên ta có thể dùng giá trị xác suất của thống kê t hoặc giá trị xác suất của thống kê F đều được. Trong trường hợp ta khảo sát nhiều hơn hai biến thì ta chỉ dùng thống kê F.

13.4. Kiểm định thừa biến trong mô hình (biến không cần thiết)

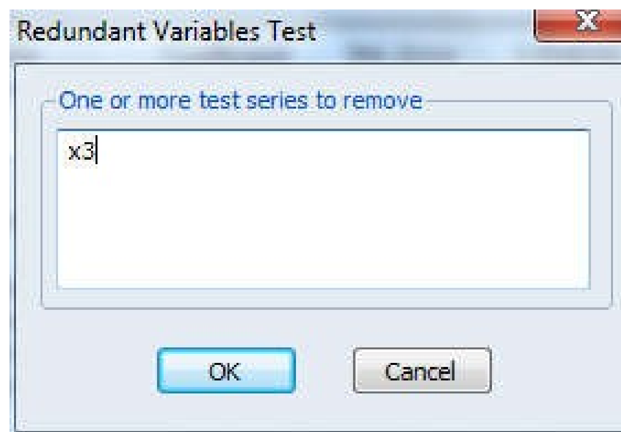
Giả sử xét ví dụ 4 bên trên, ta tiến hành như sau:

- Tìm hàm hồi quy của Y theo X2 và X3. Từ cửa sổ Equation, ta chọn **View→Coefficient Diagnostics → Redundant Variables Test – Likelihood ratio…** Khi đó màn hình sẽ như sau:



Hình 62

Nhấp chuột ta có cửa sổ **One or more test series to remove** xuất hiện, rồi gõ biến X3 vào



Hình 63 **Nhấp**

Ok, ta có kết quả sau:

Equation: UNTITLED Workfile: VIDU5::Untitled\

Redundant Variables Test
Equation: UNTITLED
Specification: Y C X2 X3
Redundant Variables: X3

	Value	df	Probability
t-statistic	11.59572	9	0.0000
F-statistic	134.4608	(1, 9)	0.0000
Likelihood ratio	33.22605	1	0.0000

Hình 64

Ta đặt bài toán kiểm định như sau: $H_0: \beta_3 = 0$: Biến X3 không cần thiết trong mô hình; $H_1: \beta_3 \neq 0$: Biến X3 cần thiết trong mô hình.

Từ bảng kiểm định ở trên, ta có $P_value = 0.0000 < \alpha$ cho trước nên bác bỏ H_0.

Vậy X3 cần thiết trong mô hình.

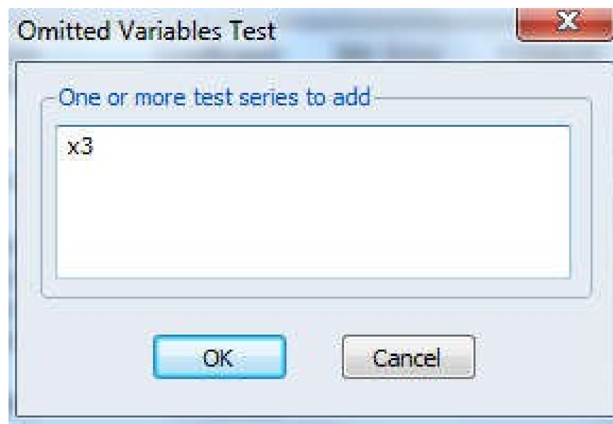
13.5. Kiểm định biến bị bỏ sót trong mô hình

Giả sử xét ví dụ 4 bên trên, ta tiến hành như sau.

- Tìm hàm hồi quy mẫu của Y theo X2. Từ cửa số **Equation**, ta chọn **View→Coefficient Diagnostics → Omitted Variables Test – Likelihood ratio…** Khi đó màn hình sẽ như sau:

Hình 65

Nhấp chuột ta có cửa sổ **One or more test series to add** xuất hiện. Ta gõ biến X3 vào



Hình 66 **Nhấp**

Ok, ta được kết quả sau:

Equation: UNTITLED Workfile: VIDU5::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Omitted Variables Test
Equation: UNTITLED
Specification: Y C X2
Omitted Variables: X3

	Value	df	Probability
t-statistic	11.59572	9	0.0000
F-statistic	134.4608	(1, 9)	0.0000
Likelihood ratio	33.22605	1	0.0000

Hình 67

Ta đặt bài toán kiểm định như sau: H_0: □ $□_3$ 0: Biến X3 ảnh hưởng tới Y (X3 không bị bỏ sót); H_1: □ $□_3$ 0: Biến X3 bị bỉ sót trong mô hình.

Từ bảng kiểm định ở trên, ta có P_ value □ 0.0000 □□ cho trước nên bác bỏ H_0.
Vậy X3 bị bỏ sót trong mô hình.

13.6. Kiểm định Chow trong mô hình hồi quy với biến giả

Ví dụ7. Giả sử số liệu về tiết kiệm và thu nhập cá nhân ở nước Anh từ năm 1946 đến 1963 (đơn vị pound) cho ở bảng sau:

Thời kỳ I	Y	X	Thời kỳ II	Y	X
1946	0.36	8.80	1955	0.59	15.5
1947	0.21	9.40	1956	0.90	16.7
1948	0.08	10.0	1957	0.95	17.7
1949	0.20	10.6	1958	0.82	18.6
1950	0.10	11.0	1959	1.04	19.7
1951	0.12	11.9	1960	1.53	21.1
1952	0.41	12.7	1961	1.94	22.8
1953	0.50	13.5	1962	1.75	23.9
1954	0.43	14.3	1963	1.99	25.2

Bảng 10

Trong đó, Y : Tiết kiệm ; X : Thu nhập.

Để kiểm định rằng có sự thay đổi về tiết kiệm giữa hai thời kỳ hay không, ta thực hiện các bước kiểm định Chow như sau:

Hồi quy Y theo X, ta được kết quả

Equation: UNTITLED Workfile: UNTITLED::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

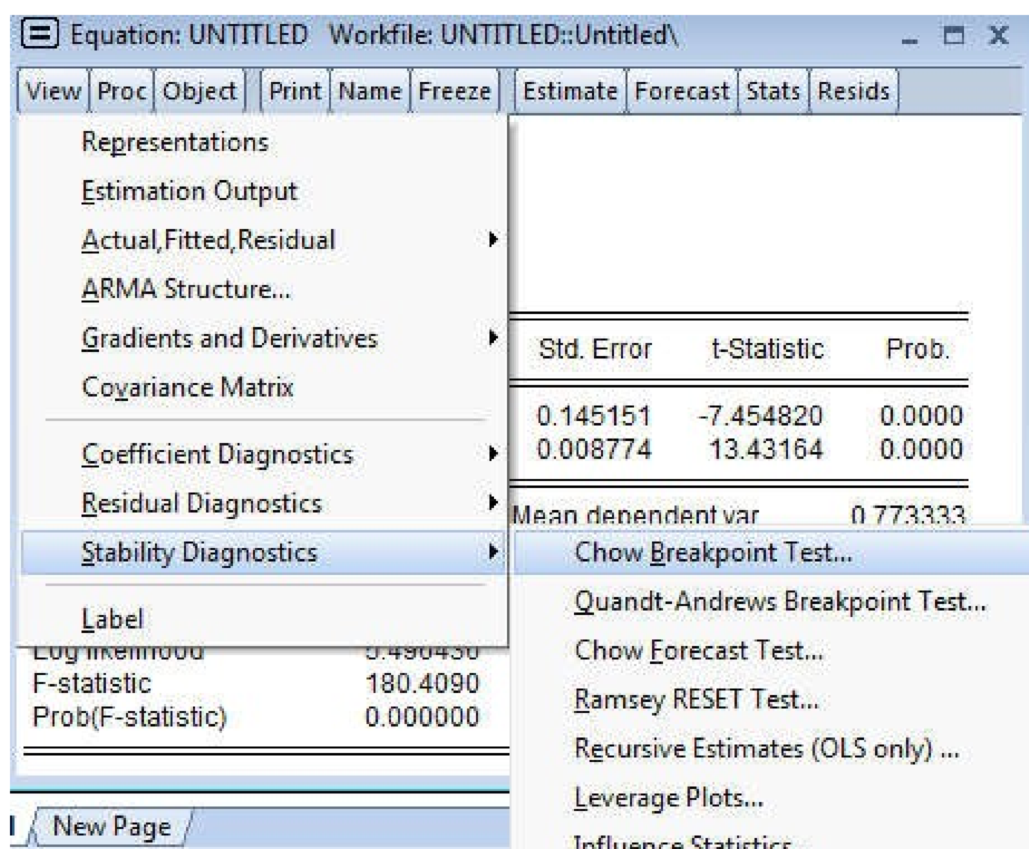
Dependent Variable: Y
Method: Least Squares
Date: 06/10/13 Time: 14:07
Sample: 1946 1963
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.082071	0.145151	-7.454820	0.0000
X	0.117845	0.008774	13.43164	0.0000

R-squared	0.918537	Mean dependent var	0.773333
Adjusted R-squared	0.913446	S.D. dependent var	0.642806
S.E. of regression	0.189114	Akaike info criterion	-0.388493
Sum squared resid	0.572226	Schwarz criterion	-0.289563
Log likelihood	5.496436	Hannan-Quinn criter.	-0.374852
F-statistic	180.4090	Durbin-Watson stat	0.947783
Prob(F-statistic)	0.000000		

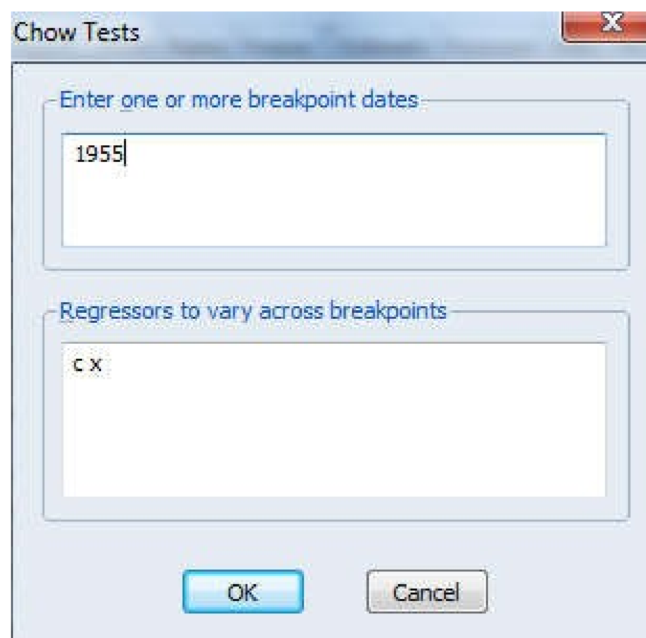
Hình 68

Từ cửa sổ **Equation,** chọn **View →Stability Diagnostics → Chow Breakpoint Test**…như hình sau:



Hình 69

Sau khi nhấp chuột, một cửa sổ xuất hiện như sau:



Hình 70

Ta gõ vào cửa sổ **Chow Test** giá trị **Breakpoint** là **1955** như hình trên, **nhấp OK**. Khi đó ta được kết quả sau:

Equation: UNTITLED   Workfile: UNTITLED::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Chow Breakpoint Test: 1955
Null Hypothesis: No breaks at specified breakpoints
Varying regressors: All equation variables
Equation Sample: 1946 1963

F-statistic	5.037060	Prob. F(2,14)	0.0225
Log likelihood ratio	9.757441	Prob. Chi-Square(2)	0.0076
Wald Statistic	10.07412	Prob. Chi-Square(2)	0.0065

Hình 71

và dự vào bảng kết quả trên ta cũng có giá trị F = 5.037. Với giá trị xác suất là 0.022493. nên ta chấp nhập giả thuyết là hai mô hình hồi quy khác nhau.

14. Định dạng mô hình (Kiểm định Ramsey RESET) Xét mô hình gốc:

$$Y_i = \beta_1 + \beta_2 X_i + U_i \quad (1)$$

Kiểm định Ramsey RESET

$$Y_i = \beta_1 + \beta_2 X_i + \alpha_1 \hat{Y}_i^2 + \alpha_2 \hat{Y}_i^3 + \ldots + \alpha_m \hat{Y}_i^{m+1} + U_i \quad (2)$$

Bài toán kiểm định

$$H_0: \alpha_1 = \alpha_2 = \ldots = \alpha_m = 0$$

$$H_1: \exists \alpha_j \neq 0, \; j = \overline{1,m}$$

H_0 : Mô hình gốc không thiếu biến, dạng hàm đúng

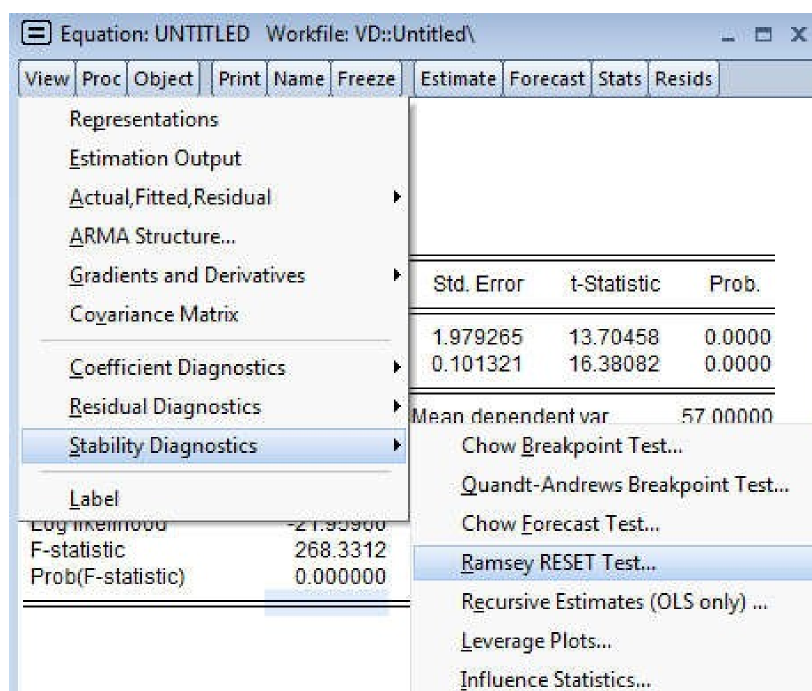
H_1 : Mô hình gốc thiếu biến, dạng hàm sai

$$F = \frac{(R_2^2 - R_1^2)/m}{(1 - R_2^2)/(n - k_2)} \sim F(m, n - k_2)$$

Giả sử xét ví dụ 3 bên trên, ta tiến hành như sau:

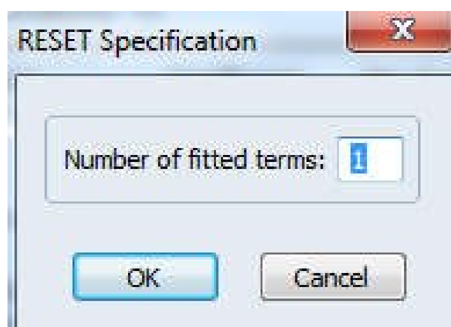
Tìm hàm hồi quy tuyến tính mẫu của Y theo X. Từ của số Equation.

Chọn **View →Stability Diagnostics →Ramsey RESET Test**…như hình sau:



Hình 72

Nhấp chuột ta có cửa sổ **Number of fitted terms** xuất hiện. Ta gõ tham số m=1 vào



Hình 73 **Nhấp**

Ok, ta được kết quả sau:

Equation: UNTITLED Workfile: VD::Untitled\

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Ramsey RESET Test
Equation: UNTITLED
Specification: Y C X
Omitted Variables: Squares of fitted values

	Value	df	Probability
t-statistic	1.177282	7	0.2776
F-statistic	1.385994	(1, 7)	0.2776
Likelihood ratio	1.806528	1	0.1789

Hình 74

Ta đặt bài toán kiểm định như sau: H $:_0$ □ □$_1$ 0: Mô hình trên không thiếu biến,

dạng hàm đúng; H :$_1$ □ □$_1$ 0: Mô hình trên thiếu biến dạng hàm sai.

Từ bảng kiểm định ở trên, ta có P_ value(F_statistic) □ 0.2776 □□ cho trước nên chấp nhận H .$_0$ Vậy mô hình trên không thiếu biến, dạng hàm đúng.

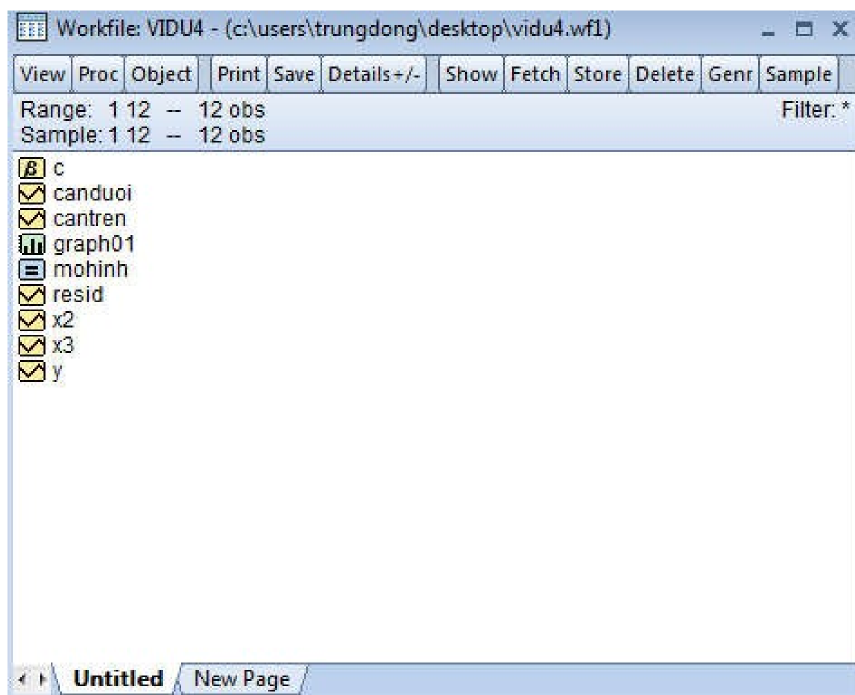
15. Lưu kết quả trong Eviews

15.1. Lưu file dữ liệu

Các thao tác được thực hiện như sau:

Sau khi làm xong các thao tác. Từ cửa sổ **Eviews** chọn **File →Save**

Lưu ý: Khi đó trên cửa sổ Workfile thì không có đối tượng nào được chọn (Nếu không ta chỉ lưu được một file dạng rác).



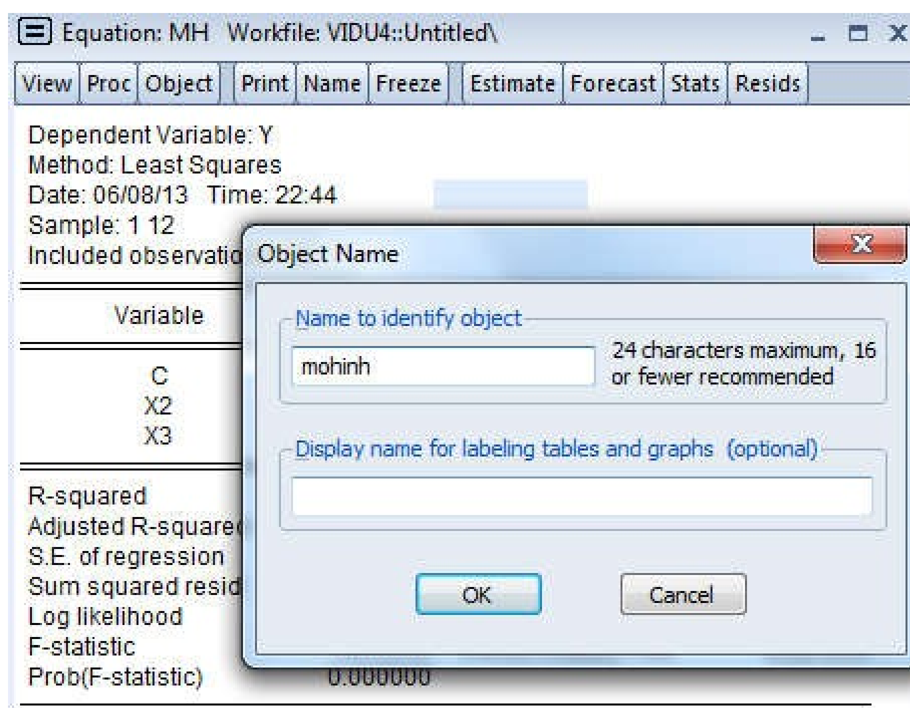
Hình 75

15.2. Lưu các bảng kết quả

Trên các cửa sổ như **Equation, Graph, Group**, …Đều có thanh công cụ chứa hai nút là : Name và Freeze dùng để lưu trữ các đối tượng hoặc các kết quả được tạo ra trong quá trình thao tác. Đối với chức năng **Name** cho phép ta lưu trữ các kết quả mà ta có thể dùng tiếp cho các thao tác sau. Mặt khác chức năng **Freeze** chỉ lưu các kết quả dưới dạng một Table (Kết quả đó được đóng băng).

Chẳng hạn với số liệu trong ví dụ 3 sau khi tìm được mô hình hồi quy xong và ta thực hiện lưu trữ như sau:

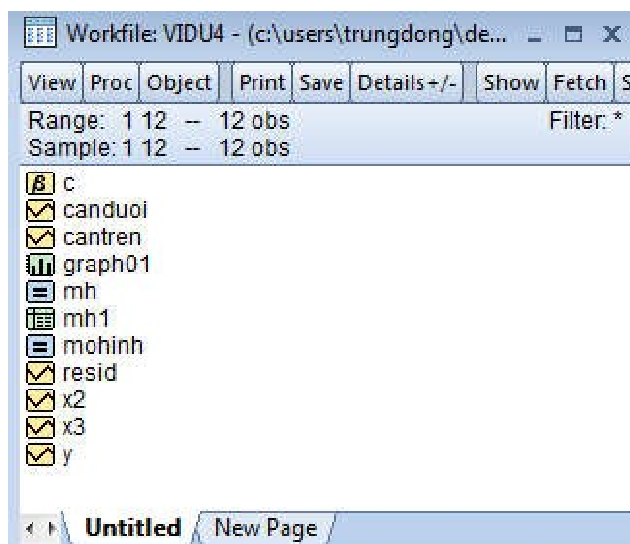
Từ cửa sổ **Equation.** Nếu ta chọn chức năng **Name** như hình 63



Hình 76

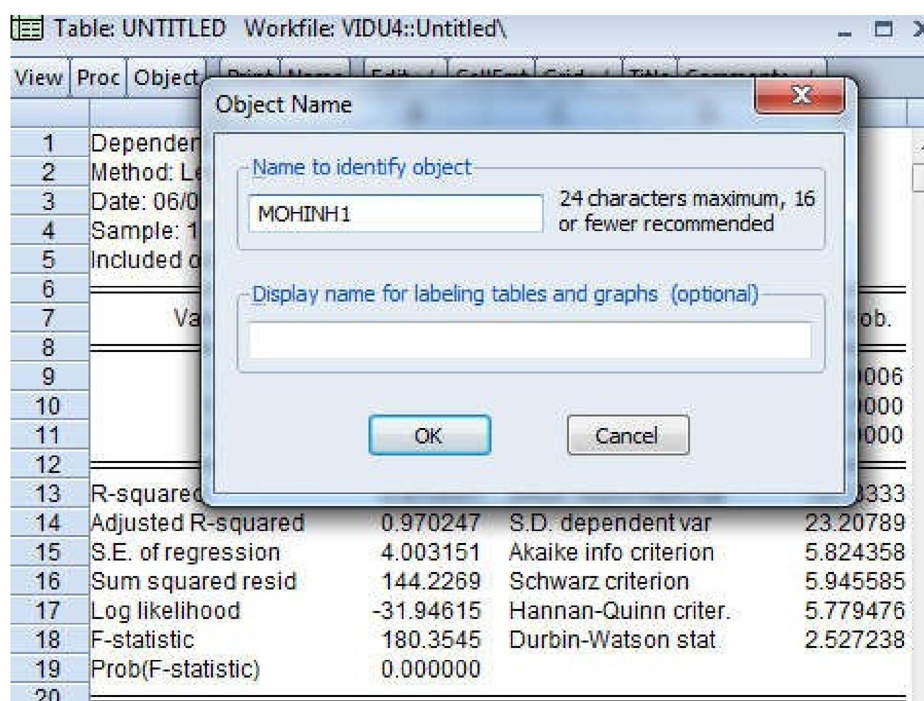
Chọn **OK** ta được kết quả có biểu tượng là





Hình 77

Từ cửa sổ **Equation**. Nếu ta chọn chức năng **Freeze** thì ta thấy một table mới xuất hiện như sau:



Hình 78

Chọn OK ta được kết quả có biểu tượng là



TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Nguyễn Quang Dong, Nguyễn Thị Minh, (2013). *Giáo trình Kinh tế lượng*, Nhà xuất bản Đại học Kinh tế Quốc Dân.

[2] Trần Tiến Khai, (2012). *Phương pháp nghiên cứu kinh tế - Kiến thức cơ bản*. Khoa Kinh tế Phát triển. Đại học Kinh Tế TP. Hồ Chí Minh. Nhà Xuất Bản Lao Động Xã Hội.

[3] Nguyễn Thị Tuyết Mai, Nguyễn Vũ Hùng, (2015). *Phương pháp điều tra khảo sát: nguyên lý và thực tiễn*, Nhà xuất bản ĐHKTQD.

[4] Hoàng Trọng – Chu Nguyễn Mộng Ngọc, (2008). *Phân tích dữ liệu nghiên cứu với SPSS*. TP.HCM. NXB Thống Kê.

[5] Nguyễn Văn Hiến, (2016), *Nghiên cứu marketing thực hành*, Nhà xuất bản tài chính.

[6] Lê Công Hoa, Nguyễn Thành Hiếu, (2014), *Nghiên cứu kinh doanh*, Nhà xuất bản Đại học Kinh tế Quốc Dân.

[7] Đinh Phi Hổ, (2016), *Phương pháp nghiên cứu kinh tế & viết luận văn thạc sĩ*, Nhà xuất bản Đại học Kinh tế TP. Hồ Chí Minh.

[8] Hồ Đăng Phúc, *Sử Dụng Phần Mềm SPSS Trong Phân Tích Số Liệu*, Nhà xuất bản Khoa học và Kỹ thuật.

[9] Nguyễn Văn Thắng, (2014), *Thực hành nghiên cứu trong Kinh tế và Quản trị kinh doanh*, Nhà xuất bản Đại học Kinh tế Quốc Dân.

[10] Nguyễn Văn Tuấn, (2011), *Cách viết đề cương nghiên cứu khoa học*. Bài giảng cho sinh viên Đại học Quốc gia TPHCM, Sydney: ĐH New South Wales.

[11] Một số bài báo quốc tế, trong nước, đề cương, bài nghiên cứu khoa học, luận án làm mẫu.

Tiếng Anh

[12] Babbie, E.R., (2011). *The Practice of Social Research*. Belmont CA: Wadsworth.

[13] Daniel Muijs, (2004). *Doing Quantitative Research in Education with SPSS*. Sage Publications.

[14] Ehrenberg, A.S.C., (1994). Theory or Well-Based Results: Which Comes First. *In Research Traditions in Marketing* (Laurent, G and Lilien, G.L.) Boston: Kluwer Academic.

[15] Kumar, R. (2014). *Research Methodology. A Step-by-Step Guide for Beginners*. Foutrth edition. SAGE Publications.

[16] Kothari, C.R., (2004). *Research Methodology: Methods and Techniques.* New Age International (p) Ltd.

[17] Marshall, C., & Rossman, G. B. (2006). *Designing Qualitative Research* (4 th ed.). Thousand Oaks, CA: Sage.

[18] John A. Sharp, John Peters and Keith Howard. (2006). *The management of a Student Research Project.* Third edition. Gower Publishing Company.

[19] Joseph F. Hair Jr. William C. Black Barry J. Babin Rolp E. Anderson, (2014). *Multivariate Data Analysis*. Pearson New International Seventh Edition.

[20] Robert A. Day. (1998). *How to write and publish a scientific paper*. Fifth edition. Oryx Press.

[21] Shuttleworth, M., (2008). Definition of Research. <https://explorable.com/definition - of - research>.

[22] Uma Sekaran and Roger Bougie (2009). *Research methods for business: A skill building approach.*